

How not to reform an independent academy

Openness is a virtue, but its imposition from without threatens the independence of the US National Academy of Sciences complex.

Over many years, the US academy complex — comprising the National Academy of Sciences, the National Academy of Engineering, the Institute of Medicine and the National Research Council (NRC) — has developed an enviable record in providing independent and penetrating advice to the federal government.

The complex, which employs some 1,100 staff, produces about 200 reports a year on a wide range of scientific and technical issues. This week's assessment of the US government's clean car programme (see page 528) is a typical example: it makes several pointed criticisms of the programme, and provides a commendably comprehensive update on all aspects of clean car technology.

The thoroughness of the academy complex's output stands in stark contrast to the voluminous material that special interest groups and government agencies themselves produce to feed the Washington policy-making machine. However, recent legal action by animal rights and environmental groups has sought to subject the work of the complex to the Federal Advisory Committee Act (FACA), one of the so-called 'sunshine laws', which requires advisory panels set up by the government to conduct all meetings in public and to adhere to stringent rules in appointing panels and conducting business (see *Nature* 385, 755; 1997 & 386, 309; 1997).

Who could object to such a requirement? The Supreme Court believed that it could not, when it cited the National Academy of Sciences as an example of a "quasi-public" body, and therefore subject to FACA, in a footnote to a 1989 ruling. But the problem with this judgement is that, in treating the academy complex as though it were part of the federal government, it ignores the very need for independence that led to that body's establishment in the first place.

The National Academy of Sciences is an élite body of 1,800 of the United States' most distinguished scientists. The foundation of the NRC in 1918, and its expansion after the Second World War, brought the academy a strong professional staff — now shared by the National

Academy of Engineering and the Institute of Medicine — and equipped its panels to draw membership from outside the élite membership. This structure allows the NRC to ensure the impartiality of the process, even when the government pays for the studies. The panels are necessarily shielded from outside pressure, particularly pressure from the federal government itself: this means that much of their discussion takes place in closed session.

The track record of the academy suggests that this arrangement succeeds in producing studies that the nation can trust. This year, for example, the Department of Energy (DoE) will ask one of its own advisory committees, operating under FACA, to assess the current design for the proposed International Thermonuclear Experimental Reactor (ITER). It will also ask the academy for its assessment. Ultimately, the Congress will place greater reliance on the independent judgement of the academy than it will on the DoE's advisory panel. On politically charged issues, such as pollution control or nuclear waste disposal, this credibility gap between the government's FACA panels and the academy's work is even more pronounced.

FACA has certainly succeeded in shining some valuable light on the workings of US government agencies. The sometimes cumbersome operation of its committees is doubtless an improvement on what went on before, behind closed doors. But its insistence on long and unproductive open sessions, and more importantly the degree of civil service supervision that it imposes, would undermine the academy complex.

The complex has recently been making efforts to keep interested parties better apprised of its work, and to open more of its meetings. Because of the autonomy of the NRC's different boards, this process has been uneven, and efforts must continue to broaden it. Such incremental reform is the best way of opening up the complex without compromising the quality of its work. If continuing court action fails to re-establish the academy's independence from the government and exempt it from FACA, the Congress must change the law to do so. □

Small symbols of peace

The Middle East provides a setting for science policy as a lever for hope amid strife.

Political events surrounding Israel give little room for any response other than despair. As the prime minister, Binyamin Netanyahu, confronts international reaction to his expansionist policy towards settlements, and as Palestinians and Israelis confront each other in spiralling violence, one looks for signs of hope that rigidly opposed viewpoints can be softened and mutual understanding enhanced.

All avenues need to be pursued, and science provides one. Cooperation in research between Israel and its neighbours has been developing in recent years. Israel's Arab neighbours have not always been able or willing to rush to participate in joint research, while Israel itself is not a development agency. A reluctance to engage, and the disparity in scientific resources and achievement, is

especially marked with the Palestinians.

In such circumstances, a third party with money to spend can play a useful role. One such is Germany's funding agency, the Deutsche Forschungsgemeinschaft (DFG). It is funding German applicants in collaborative programmes with Israel on condition that Palestinians are also involved. Four projects are under way — in cancer genetics, developmental biology, agriculture and geographic information systems, with more to come. Similar trilateral schemes are funded by Germany's research ministry and by the United States.

One must hope that such projects will quietly prosper despite the circumstances. Agencies elsewhere with scope to tailor foreign collaborative schemes to boost reconciliation, in the Middle East or elsewhere, should consider following the DFG's example. □



'EU programme should increase support for social sciences'

[LISBON] The social sciences should play a more prominent role in plans for the European Union's fifth Framework programme for research (FP5), due to start in 1999, according to the research ministers of a number of member states.

Meeting in Lisbon, Portugal, on 4–6 April, ministers from Denmark, the Netherlands, Sweden, Portugal and Italy — as well as non-EU member Norway — gave their broad agreement to the idea that more needs to be and could be done to promote the social sciences at European level.

They suggested that the social sciences would benefit from more support from Brussels for the type of strategies that have been applied successfully to the natural sciences — such as the creation of centres of excellence, the expansion of Europe-wide databases, and enhanced networking.

But some ministers, such as Jo Ritzen, the minister of education, science and culture in the Netherlands, went further to suggest that the growing economic and political integration of Europe is creating a new form of society, which the social sciences should take as an increasing focus of their research.

The meeting was organized by José Mariano Gago, Portugal's minister of science and technology, who said that it had arisen from a growing feeling that the social sciences need a higher profile in European science policy. "The political construction of Europe is an enormous challenge for the social sciences, just as it has been to natural scientists," said

Gago, who was formerly a physicist at the European Laboratory for Particle Physics (CERN) in Geneva, Switzerland.

Edith Cresson, European commissioner for research, said that current draft proposals for the four-year FP5 programme, which were due to be presented to the full commission this week (see box), already contain a strong commitment to support research in the social and economic sciences. The proposals aim in particular at helping to meet identified social goals, such as future transport systems.

"The social sciences have to give us a better capability to handle the consequences of the introduction and dissemination of new technologies, and they can also guide our research in the innovation field and improve our ability to target specific social needs," said Cresson.

Ironically, it is precisely such efforts to introduce a social agenda into technical research programmes that have upset some of the larger — and politically more conservative — EU member states. Their governments have been demanding a narrower focus to the Brussels research agenda, and were notable by their absence from the meeting in Lisbon.

But several participants argued that it was Europe's weakness in fields such as management science — as much as in the natural sciences — that is responsible for the relative weakness of its industrial economy compared to that of the United States.

A background paper produced by Italy's research ministry, for example, argued that



Cresson: planning for 'social agenda'.

there is "growing recognition in Europe that the weakness of the system lies not in the scientific ideas, but in the process of transforming them into useful and beneficial ideas for mankind".

The focus of discussion during last weekend's meeting was not whether, but how, enhanced support should be established for the social sciences at European level. Carl Tham, Sweden's minister for education and science, argued that, beyond supporting socio-economic research ties to specific objectives, FP5 should include a specific programme or 'key action' on the social sciences and humanities.

Guido Martinotti, former chair of the social science committee of the European Science Foundation, urged that a broader range of social sciences should be brought to bear on priorities that have already been identified, such as 'the city of tomorrow'. This view reflected complaints of an excessive emphasis on the technical dimension of such issues.

Politically, the impact of these discussions on the detailed contents of FP5 could depend to some extent on whether the Labour Party wins the British general election — and, if so, how far it is prepared to support the predominantly social democrat governments now seeking greater financial backing for the social sciences from Brussels.

David Dickson

Fellowships seem likely to win out over basic research in Europe

[MUNICH] The European Union's fifth Framework research programme (FP5) looks set to give much more support to research fellowships, but non-targeted basic research projects will not be a priority.

The European Commission's formal proposal for the five-year programme is expected to be made public this week. It will then need to win the approval of member states and the European parliament.

The proposal asks for the total budget of FP5 to be kept at least the same level, relative to member states' gross national product, as the ECU13.2 billion (US\$15.2 billion) of the fourth Framework programme (FP4). But it does not put an exact price

on FP5, since the commission's overall budget for its running period — 1999 to 2003 — has not yet been agreed. It is likely to be between ECU15 billion and ECU17 billion, according to observers.

The proposal holds few surprises for those who have seen early drafts, which were widely circulated outside the commission as part of its new policy of consulting a range of interested parties (see *Nature* 386, 5; 1997). It suggests three thematic programmes covering life sciences and the environment, information sciences, and industrial technologies. Each is earmarked for 28 per cent of the total funds.

Each thematic programme is divided into a series of strategic key actions plus an element dedicated to non-targeted generic technologies and support for infrastructure. The element for non-targeted basic research has disappeared. This was highlighted in earlier drafts, but was criticized by several member states, including Germany and Britain (see *Nature* 386, 205; 1997). Opponents had argued that basic research should be carried out only in the areas defined in the key actions.

The three additional, so-called 'horizontal', programmes receive different funding from that of comparable programmes in FP4. The budget share going to

international cooperation falls from 5 per cent to 3.5 per cent, while that going to dissemination of research results remains at 2.5 per cent.

But the third horizontal action — training and mobility of researchers — is to increase its share from 6.5 to 10 per cent. It had become clear to the commission in the discussions with member states that this programme, which supports young European researchers working in a second EU country, was universally popular.

Euratom, Europe's atomic energy community, sees its share of the Framework 'cake' drop from 10 to 9 per cent.

Allison Abbott

Clean car programme falling short of goal

[WASHINGTON] One of the Clinton administration's flagship technology programmes will not meet its main goal of developing by 2004 a competitively priced, family-sized 'clean car' that does 80 miles to the gallon, according to a National Research Council (NRC) expert panel.

The panel says the Partnership for a New Generation of Vehicles (PNGV), bringing together the federal government and the 'big three' US car-makers — General Motors, Ford and Chrysler — will need more money and a sharper focus on research and development problems if it is to meet the target.

Substantial progress has been made on several key technologies, such as fuel cells and lithium-ion batteries, says the NRC panel. But other major problems remain unresolved, such as the development of flywheels or large capacitors for energy storage.

As a result, the partnership's plan to select viable technologies this year for the car to be built in 2004 is "untenable", says the panel. That would force the programme to support only the most conservative options — such as diesel engines — and to discard the more difficult technologies which hold greater long-term promise.

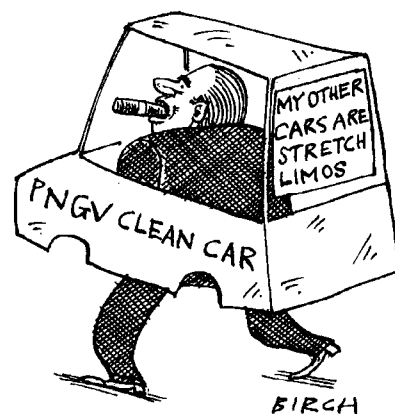
The panel also says that the programme is continuing to put money into technologies, such as large capacitors, that have little

potential to help PNGV meet its goals. And it has neglected vital problems that lack appeal to engineers, such as reducing the power consumption of car heating, cooling and lighting systems.

Trevor Jones, chairman of the NRC panel and of Echlin, a major US car parts supplier, says PNGV "should focus research on these areas that need major breakthroughs". He says the Clinton administration has done a good job of coordinating the programme, but has had trouble with the Congress in transferring money between elements of it as needed. PNGV has a budget this year of \$270 million, spread across four government departments and several smaller agencies.

In a joint statement, the government and its three partners said that they were "pleased that the review recognised the significant progress PNGV has made" in the past year. "The formidable challenges were known from the start," the statement said. "The partners remain enthusiastic about the prospects of the program and its ultimate benefits to the nation."

Ron York, PNGV director at General Motors, says that the programme is likely to adjust its technology selection process to ensure that work continues on items of long-term potential. He says that "the industry side will be recommending to the govern-



ment that work continues on these" as the car manufacturers move resources towards building prototypes.

"It's very hard to communicate how hard this kind of work is," says York. "We didn't get to the moon in one giant leap. If this was easy, we'd have done it a long time ago."

Some Republicans in Congress were initially critical of PNGV, which they saw as "corporate welfare", but the car-makers have had some success in convincing them of its value, and last year it was budgeted at a steady level. But the programme does not look well-placed to obtain the unspecified additional funding that the NRC report calls for.

Colin MacLachlan

US supercomputer centres to concentrate on networking

[WASHINGTON] Two distributed supercomputer centres to be supported by the National Science Foundation (NSF) will give US university researchers quicker access to the most advanced supercomputer systems than has been possible with the existing four centres, NSF officials have promised.

The directors of the two centres, at the University of California at San Diego and the University of Illinois at Urbana-Champaign, pledged a 'seamless' transition for the users of the two other centres, which will now lose NSF support, and promised that their new annual budgets of around \$30 million will allow more frequent hardware upgrades. Both centres also plan to help their users tap into far larger supercomputers being installed at the US nuclear weapons laboratories.

According to Bob Borchers of the NSF's computing directorate, the new centres will spend about one-third of their money on hardware, one-third on operations at their headquarters, and the rest on operations at their distributed partners' sites.

Larry Smarr, director of the Illinois centre, says the idea for its distributed approach — linking the processing power of

supercomputers at different centres — arose when an experimental network was set up between 50 universities and laboratories for a supercomputing trade show in 1995. "That was such a phenomenal thing," he says. "For three days, we were living as though it was the 21st century."

The Illinois centre, which will lead a partnership called the National Computational Science Alliance, will use computers built by Cray Research (now part of Silicon Graphics) and plans to concentrate on creating a high-speed network, linking together its dozens of university and industrial partners.

The San Diego centre will use IBM hardware, and is already negotiating a deal with the Lawrence Livermore National Laboratory to access part of an extremely powerful IBM machine being installed there for nuclear weapons research.

But managers of the centres that will now lose funding, at the University of Pittsburgh, Pennsylvania, and Cornell University, New York, are complaining bitterly about the NSF's selection process, and warn that their users could face a difficult transition. Their complaints will be aired this week at a hearing of the basic

science subcommittee of the House of Representatives' Science Committee.

Cornell unsuccessfully challenged the selection process in February. "I have concerns that the process had a fair number of flaws in it," says Norman Scott, vice-president for research at Cornell. "It wasn't up to the high standards we normally subscribe to the National Science Foundation."

The process that led to the selection of San Diego and Illinois by the National Science Board on 28 March was dogged by the disqualification of many reviewers and board members who were associated with one of the dozens of collaborating institutions involved in the four rival bids.

Scott also says many users are sceptical about the prospects for a smooth transition to the new centres, and that the small amount of NSF funding — \$11 million over two years — for Cornell to wind down its operations has "surprised us dramatically".

But Borchers says the two winners were "qualitatively better, and highly complementary to each other. They went beyond what we thought the vision would be, and offered a broader approach which was really rather compelling".

C.M.

Creationism 'Ark' trial opens in Australia

[SYDNEY] The running battle between creationists, scientists and educators over the teaching of evolution in schools took a new turn this week with the opening of a trial in Sydney, Australia, in which a leading geologist is claiming that the public has been deliberately misled by statements about what some are claiming to be the petrified remains of Noah's Ark.

The creationist dispute has been raging ever since the infamous 'monkey trial' in Tennessee 72 years ago, at which John Scopes was convicted for teaching evolution. The Australian case represents the culmination of highly publicized confrontations that took place in 1992 between Ian Plimer, head of Earth sciences at the University of Melbourne, and Allen Roberts, a proponent of creationism.

Plimer says that Roberts made false claims at lectures about 'Noah's Ark' — a boat-shaped structure found at Akyayla in the mountains of eastern Turkey, 20 km southeast of Mount Ararat. Plimer wants the federal court to restrain Roberts from repeating his claims and to stop sales of tapes of the lectures. The tapes were being sold by Roberts's support group, Ark Search Association Incorporated.

The case is the first time that the relative credibility of science and creationism have been disputed under legislation designed to protect consumers against misleading or deceptive commercial conduct. The thrust of Plimer's case lies in the legal responsibility of a trader not to make false or unprovable claims.

Plimer says he has been motivated to pursue Roberts partly by the results of surveys of new students in the science faculties of Australian universities which show that up to one in five holds creationist views. Australia has a growing number of private schools run on fundamentalist Christian principles, and these have criticized Plimer,

on the grounds of supporting freedom of speech, for not agreeing with their teaching alternative 'theories' of life.

Roberts claims that he and associates found 'scientific' evidence of Noah's Ark in the boat-shaped structure. It has been known for at least 40 years that the structure exhibits the dimensions of the Ark as recorded in Genesis. In lectures, publications and tapes sold to support an expedition for further 'investigation', Roberts has strongly suggested that verification — he has never talked of 'proof' — of the structure's origins came from the presence of the ribs of a boat, petrified wood, iron rivets, a deer's antlers, fossilized animal dung (coprolites) and stone anchors.

During a contentious lecture tour by Roberts in 1992, police were called to eject Plimer and other scientists seeking to challenge his statements. Roberts was described by Ark Search as its "archaeological research consultant".

Rather than being asked to decide on the relative claims of science and creationism about the nature of the Turkish structure, Judge Ron Sackville will assess claims made by Plimer and David Fasold, a US marine salvager, against Roberts and Ark Search on commercial grounds. The case is a civil trial, with applicants and respondents arguing their case before the judge, but without a jury. In early hearings before witnesses were called, the judge indicated that he would look closely at assertions that "activities" at the site by either Plimer or Roberts were "scientific in character".

Fasold was a firm believer in the existence of the Ark, and published a book describing his views in 1988. He alleges that Roberts infringed his copyright, as a result of which he "suffered loss and damage". He now says the structure is not the Ark, but what the ancient residents of the Euphrates Valley thought to be so.

On one of several field trips to Turkey, Plimer studied the Akyayla site in 1994. He concluded that it is an unremarkable formation of folded and weathered ophiolite, partly covered in mud, and suggested that it is between 110 million and 120 million years old. In a book published that year, he accused creationists of "scientific fraud" in promoting the site as the Ark.

Plimer suggests that Roberts and Ark Search have made misleading claims about their involvement in gathering data about the 'Ark' in a scientific manner. Plimer also challenges the way in which Roberts justified his claims to scientific expertise by citing a doctorate in Christian education which, claims Plimer, "is from an institution of low academic repute and has no relevance to archaeological or scientific knowledge".

Roberts describes himself as a graduate of Freedom University, a correspondence Bible college which was revealed in 1992 to have little presence beyond sharing a mailbox with the Maitland Bible Church in Orlando, Florida.

Roberts and Ark Search deny all the claims and damages. As part of their defence, they claim Fasold "formed an association" with Ron Wyatt, with whom he "cooperated and shared information" in studying the Ark site. Wyatt, from Nashville, Tennessee, described himself in 1989 as having been officially recognized by the Turkish government as "the archaeologist who discovered Noah's Ark".

On the advice of their lawyers, Roberts and Ark Search have declined to provide interviews or statements "before or during the trial".

Fasold and Plimer are calling four witnesses. Two are Australian palaeontologists, Alex Ritchie of the Australian Museum in Sydney and Neil Archbold of Deakin University in Melbourne. One, Edwin Byford, is a theologian, and the fourth, Eugenie Scott, is director of the National Center for Science Education in Berkeley, California, and a long-term campaigner against the teaching of 'creation science' in American schools.

The trial is scheduled to last three weeks, and the costs to the loser will be high. Plimer estimates that the case is already costing him A\$500,000 (US\$390,000), for which he has had to sell his house.

Six days before the trial, Ark Search went into liquidation, complicating an already difficult case for Plimer to argue within commercial law.

The outcome of the trial is likely to have a direct impact on another case in which Plimer is being sued by Roberts in the Supreme Court of Victoria for alleged defamation in public and in the media.

Peter Pockley



Noah's Ark? Claims about this geological formation near Mount Ararat, Turkey, form part of the case.

Biologists sound warning on species conservation

[WASHINGTON] A small but influential group of conservation biologists last week warned that efforts to make the US Endangered Species Act (ESA) more friendly to property owners "will not further the Act's goals unless those measures are implemented in a scientifically sound manner".

The ESA lets landowners develop land that includes protected habitats provided they develop Habitat Conservation Plans (HCPs) which mitigate the harm this causes protected species by setting aside habitats or making other actions to protect species in the long term. But better scientific review of such plans is needed, warns an ad-hoc group of nine biologists and geneticists led by Dennis Murphy of Stanford University. The group sent its report last week to Katie McGinty, head of the White House Council on Environmental Quality, and congressional environmental committee chairmen.

The use of HCPs has proliferated during the Clinton administration, which has tried to ease political tensions between landowners and environmentalists. But the increase in the scale of the plans, many of which cover

many species and enormous areas of land over long periods, has concerned environmental scientists. "Many recent HCPs have been developed without adequate scientific guidance," write the biologists. They say the plans have the potential to contribute to, rather than alleviate, threats to listed species and their habitats.

The group calls for the creation of a standing body of independent scientists to review large and complex HCPs. The comprehensiveness of such reviews, they write, should be "appropriate to the size and duration of the plan". Any multiple-species HCPs must also include adequate research and monitoring programmes, with an "assured source of funds".

To be scientifically credible, the plans must "clearly articulate measurable biological goals and demonstrate how those goals will be attained". HCPs should not undermine the recovery of listed or vulnerable species by simply preserving their endangered status on isolated parcels of land.

The biologists also recommend increased funding and improved training for govern-



At risk: the Coastal Californian Gnatcatcher is one endangered species that could be affected.

ment officials in charge of enforcing the ESA, principally at the Fish and Wildlife Service and the National Marine Fisheries Service.

Further concerns have been raised by assurances given by Interior Secretary Bruce Babbitt in 1994 that landowners would have "no surprises" in complying with the Act, meaning that, once the government had approved an HCP, it would not impose additional requirements on the property owner. The biologists argue that habitats and species will be lost unless conservation plans can be amended.

The biologists agree that, "with essential stipulations, landowner-friendly initiatives can assist in meeting [the ESA's] goals". But they prefer to replace the 'no surprises' policy with a 'fair assurances' policy, which would allow an HCP to be changed based on new scientific information. But the scientists concede that the public, rather than landowners, might reasonably be asked to pay any costs associated with amending plans after they have been approved.

The group of biologists includes Peter Brussard of the University of Nevada, Michael Soule of the University of California at Santa Cruz, Reed Noss of Oregon State University, and Gary Meffe of the Department of Energy's Savannah River Ecology Laboratory. Murphy, currently president of the Society for Conservation Biology, invited them to Stanford in February to hammer out a 'consensus document' that could help add scientific objectivity and moderation to an increasingly polarized debate.

The ESA is up for review, but few are optimistic that this will happen quickly or easily. A bipartisan bill penned by two Senators on opposite sides of the issue — Dirk Kempthorne (Republican, Idaho) and John Chafee (Republican, Rhode Island) — has been in 'discussion draft' since early this year, with no apparent agreement between environmentalists and property rights advocates. The 'no surprises' policy comes up for a 60-day public comment period this spring.

Murphy invited only scientists who had worked on large conservation plans involving private property to join the group. He himself recently led a scientific review panel advising on a large and controversial HCP in southern California.

Tony Reichhardt

Discoverer of comet bemoans lack of jobs

Alan Hale, the co-discoverer of the bright comet appearing in the sky this spring, is using his new-found celebrity not to talk about the wonders of the Universe, but to call attention to the difficulties facing younger scientists. In press interviews and in an open letter on the Internet, Hale, a 38-year-old astronomer, denounces the "abysmal" opportunities facing scientists of his generation.

"Unless there are some pretty drastic changes in the way that our society approaches science and treats those of us who have devoted our lives to making some of our own contributions, there is no way that I can, with a clear conscience, encourage present-day students to pursue a career in science," he wrote in his letter, which went out to various electronic newsgroups in late March.

Since discovering the comet named after him and amateur astronomer Thomas Bopp in 1995, Hale has used his many interviews in the media to highlight the difficulties that young scientists with PhDs face in finding well-paid jobs. Hale says that reporters often do not use this part of the interview. But he says he intends to continue trying to raise awareness.

Hale searched in vain for a university position several years ago. He eventually

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Hale: 'abysmal' situation means that he cannot encourage students to pursue a career in science.

took a low-paid job in a science museum in New Mexico before creating his own research and education organization, the Southwest Institute for Space Research, in 1993.

He says he knows that others have experienced similar problems in "pursuing any kind of decent career within science". A request for "horror stories" from other young scientists has yielded hundreds of responses, he says.

T.R.

'No consensus' on FDA role in gene tests

[WASHINGTON] A US government task force has dropped plans to call for the Food and Drug Administration (FDA) to extend its regulation of genetic tests. It has decided instead not to pronounce on the issue in its final report, due to be published next month.

Neil Holtzman, director of genetics and public policy studies at the Johns Hopkins Medical Institutions, who is chairman of the Task Force on Genetic Testing, said last week that the 15-member group "is not going to recommend that FDA extend its authority to assessing genetic-test services".

Holtzman said there was no point in pursuing the issue further because of a lack of "consensus" on the matter between members of the panel, which is made up of scientists, ethicists, industry and consumer representatives.

In a closed session last month, a small majority of the group voted against calling for increased FDA involvement in regulating genetic tests.

At present, the FDA monitors genetic tests that are developed and sold as kits through doctors and other providers, but does not monitor those marketed by companies as in-house 'services'. No genetic-test 'kits' as such are known to exist; instead,

companies have developed tests as services, thereby avoiding FDA surveillance.

In provisional recommendations published in the *Federal Register* in February, the task force said that one option would be for the FDA to monitor even those genetic tests marketed as services, requiring pre-market FDA approval when such tests were considered to require particularly stringent scrutiny (see *Nature* 385, 477; 1997 & 384, 101; 1996).

Tests would be considered to fall into this category on the basis of criteria such as their ability to predict serious future disease in healthy individuals and their offspring.

But more than 50 comments received since February in response to the provisional recommendations appear to have swayed the task force against calling for FDA involvement.

The American Society of Human Genetics called the approach "heavy handed" and without "clear rationale", while the College of American Pathologists said the additional mandate would "substantially compromise" the FDA's ability to carry out its current work. The American Medical Association called the suggestion "ill advised".

According to Patricia Murphy, a repre-

sentative of the genetic-test developer OncorMed, who is a member of the task force, "Essentially half of [the task force] strongly believe [FDA] is not the right agency. They don't have geneticists, they don't have the experience, they don't have the money or the resources and they don't have the interest." She added that the Centers for Disease Control and Prevention were already well placed to carry out such work.

In other final recommendations, the task force will urge that commercial and private developers of genetic tests should be required to submit test-development protocols to local ethics panels called Institutional Review Boards, or equivalent bodies. Tests would also have to be approved by external review bodies before being used in clinical practice. Separately, both the authority and the extent of its jurisdiction of a suggested National Genetics Board would be scaled back from original proposals.

The task force was established by the Working Group on the Ethical, Legal and Social Implications of Human Genome Research of the National Institutes of Health and the Department of Energy, to make recommendations for the safe and effective development and use of genetic tests.

Meredith Wadman

Nirex chief expected to lose seat on UK's nuclear waste panel

[LONDON] Michael Folger, chief executive of Nirex, the company responsible for disposing of Britain's nuclear waste, is expected to lose his place as a member of the government's Radioactive Waste Management Advisory Committee (RWMAC).

A routine internal review by the Department of the Environment (DoE) is to recommend that Folger's term of office is not renewed. He has served on the committee since 1991. A source close to the DoE confirms that Folger's position is "under consideration". A final decision, however, rests with the new Secretary of State for the Environment, who will take office after the general election on 1 May.

A spokesman for Nirex says the company will not comment on what it considers to be "speculation". But the decision, if taken, would be yet another blow for Nirex. The company has effectively been forced to abandon plans to bury intermediate-level nuclear waste in a planned deep repository in the northwest of England since the environment secretary refused to allow the company to build an experimental research facility to test the site geology.

A decision not to renew Folger's term will not, however, be entirely unexpected.

Observers such as environmentalist groups have often said that Folger's presence on RWMAC could be seen to compromise the committee's independence, particularly when RWMAC is called upon to review Nirex's own proposals, such as its plans to build the experimental rock laboratory, or more recently, the company's peer-review arrangements. One source says Folger's presence on RWMAC gave the impression that "Nirex were reviewing themselves".

RWMAC is made up of experts and lay members. Its members are drawn from areas including the nuclear industry, universities, local government, an environmentalist group and the National Health Service.

News of Folger's possible removal has provoked a mixed reaction. The environmentalist group Friends of the Earth believes the move is long overdue. Rachel Western, senior nuclear research officer at Friends of the Earth, says, "it is essential that the government now receives advice from a broader-based community of environmentalists and scientists".

Current and former members of RWMAC, on the other hand, remain divided. Some consider the DoE's recommendation to be an overreaction. "I

can't see a problem," says one former member. "RWMAC is an independent committee, and we didn't feel gagged at all."

Other RWMAC sources believe that Folger's position was difficult to justify, given that the committee does not include relevant experts from the anti-nuclear environmentalist movement. "RWMAC should either have representatives from all sides, or be composed entirely of independent academics. If environmentalist groups are excluded, the same should apply for the nuclear industry," said one.

A DoE spokesman says that RWMAC members are chosen for their professional expertise rather than their affiliation. But one critic points out that Folger, a former civil servant with a background in finance, did not appear to be the ideal choice to represent professional expertise in nuclear waste management.

Meanwhile, Nirex has denied suggestions that it is planning to make redundant a large number of scientists. At a board meeting earlier this month, the company decided not to appeal against the government's decision to refuse its application to build the rock laboratory, presaging speculation that redundancies were likely (see *Nature* 386, 423; 1997).

Hsian Masood

ESA agrees full re-fly for Cluster mission

[MUNICH] The European Space Agency (ESA) last week decided to re-fly all four satellites of its Cluster mission. The original mission was destroyed when the Ariane V launcher exploded on its maiden flight in June last year (see *Nature* 381, 541; 1996).

Although the decision will strain ESA's already stretched budgets even further, its Science Programme Committee (SPC) felt this was justified by the scientific value of Cluster, which will measure the interactions between the solar wind and the Earth's magnetosphere.

The decision also represents recognition of the substantial development costs that ESA had invested in the original Cluster mission. But perhaps the most important factor was the relatively low cost of the relaunch. ESA has managed to keep the costs of Cluster 2 below the ceiling established for the rescue programme last year, chiefly by tough negotiations with contractors, but also by using a relatively cheap launcher, the Russian Soyuz.

The final price of ECU214 million (US \$300 million) also includes a significant contribution — around 40 per cent — from the agency towards the costs of rebuilding the scientific instruments, something normally shouldered entirely by participating member states. France, Germany and Britain, the three member states with the greatest interests in Cluster, had previously said they could not afford the full costs of



Twin peek: Cluster's satellites will get a second chance.

replacing the instruments, estimated to be around ECU40 million.

The German aerospace company Dornier will be the prime contractor for refurbishing the original Cluster mission's spare satellite — nicknamed Phoenix for the hope it gave that Cluster could rise from its own ashes — and for building three identical satellites. Cluster 2 will be launched in two pairs, on Soyuz launchers, in mid-2000.

The launch of all four satellites, rather than Phoenix alone, means Cluster will retain its unique ability to provide a three-

dimensional analysis of the magnetosphere. But according to Giacomo Cavallo, head of science planning at ESA, some of Cluster's other aims may have to be adjusted.

The four-year delay means that Cluster 2 cannot be coordinated with other satellites that are now flying, such as the solar observatory SOHO run jointly by ESA and the US National Aeronautics and Space Administration, which may end its operations in 1998. But it does mean that it will be around when sunspot activity is at a peak, in 2001 and 2002.

Alison Abbott

Europe outlines labelling rules for genetic modification seeds

[MUNICH] Consumer and environmental organizations have gained some ground in the continuing controversy over the labelling of genetically-engineered products. Last week the European Commission approved an amendment to its directive on the deliberate release of genetically modified organisms to make it compulsory for genetically engineered seeds to be labelled as such.

But the new rule will not apply retrospectively. It will therefore not affect the dozen or so modified seeds already approved for sale in the European Union, including the controversial pest-resistant maize produced by Ciba (see *Nature* 384, 503; 1996). Nor will it necessarily apply to crops that have been grown from modified seeds or to processed foods that are based on such crops.

Greenpeace is unhappy with the new rule, which it argues is not tight enough, particularly because mixtures of modified and non-modified seeds would only need to carry a label saying the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Heights of protest? Greenpeace attached banners to the towers of Cologne cathedral.

mixture 'may contain' genetically modified organisms. More generally Greenpeace is unhappy about the gap between labelling at the agricultural and consumer levels. If tomato seeds were labelled as containing genetically modified organisms, this would not mean that tomato paste made from the products of such seeds would have to be similarly labelled, says a spokeswoman.

The commission's environment directorate sees the new rule as an interim measure, and is carrying out a complete review of the directive on deliberate release of genetically modified organisms. It intends to report to the Council of Ministers and the European Parliament in the next months on the comprehensive labelling of such organisms and their products.

The only other existing directive on genetically modified organisms is that concerning novel foods. This requires labelling of genetically modified foods only if they can be shown to have characteristics that clearly differ from their non-modified counterparts.

The environment directorate would like to see much tougher rules on labelling, and would like all steps in the food chain — from production on the farm to retail sales from the supermarket shelf — to be regulated. In contrast, the trade and industry directorates believe this to be too restrictive.

A. A.

US privacy laws may curb access to medical data

[WASHINGTON] Moves in the United States to legislate on medical privacy are creating a heated debate between researchers and privacy advocates about the conditions under which scientists should have access to medical records and archives of tissue samples. At stake is the relatively liberal access to such material that researchers currently enjoy.

Under the Health Portability and Accountability Act of 1996, Donna Shalala, Health and Human Services Secretary, has until August to make "detailed recommendations" on medical privacy to Congress. If Congress does not produce legislation based on these recommendations by August 1999, Shalala must issue regulations, although their scope would be narrower, being restricted to financial and administrative transactions.

At present, medical privacy is not governed by any comprehensive federal law, although specific laws apply to some federal government programmes. But state laws on privacy, and the use of genetic information, are proliferating. A uniform federal law would bring greater coherence, argue some scientists.

A subcommittee of President Bill Clinton's National Bioethics Advisory Commission is also expected later this year to make recommendations on the use of the tens of millions of archived tissue samples, collected from autopsies, biopsies and surgery across the United States. The outcome of the debate is expected to be new federal legislation.

At present, nominative tissue samples can be used by researchers if they can persuade local ethics committees — called institutional review boards — that nominative data is essential to a study, and that its use will not harm subjects.

Many scientists argue that such access is critical to much beneficial research. Oncogenes that could be used as diagnostic or prognostic markers might be identified from stored tumour samples, for example, but that requires access to patient records showing disease outcome. The link between vaginal cancer in young women and use of diethylstilbestrol by their mothers was also established through identifiable archived specimens.

But privacy advocates argue that the risk to patients of sensitive information falling into the wrong hands outweighs the interest of scientists. James Pyles of the Coalition for Patient Rights recently told the subcommittee of the National Committee on Vital and Health Statistics that is advising Shalala on medical privacy that "simply because someone can think of a reason why they need identifiable health information to provide better health care doesn't mean they have a right to it or should be able to get it without consent".

Robert Gellman, a medical privacy consultant who is chair of the subcommittee, says: "It is an open question whether researchers will be able to get access to identifiable records at all — and even in some proposals non-identifiable [records]."

An expert panel convened by the National Heart, Lung and Blood Institute is this month expected to recommend that use of identifiable samples in an institute repository of 2.3 million blood samples requires informed consent, unless the research concerns the blood safety issues for which the samples were collected.

But a statement issued last year by 17 scientific groups convened by the College of American Pathologists argues that nominative samples should be available without consent for research approved by institutional review boards, provided effective security policies are in place. Excessive restrictions risk putting "tremendous roadblocks" in front of medical research, says Mark Sobel, the chief of the Molecular Pathology Section at the National Cancer Institute, who helped draft the statement.

A consensus is emerging among some ethicists and scientists, however, that additional protection is needed in genetics research, and that informed consent should be required for use of nominative tissue samples. Dispensations would be given only when stringent conditions were met.

Several bills being introduced in Congress may help define the spectrum of opinion there. In the House of Representatives, Gary Condit (Democrat, California) has introduced a health information privacy bill that would give researchers access to identifiable medical records without consent, provided that studies are considered by an institutional

review board to be "of sufficient importance so as to outweigh the intrusion into the privacy of the individual". A bill expected to be introduced this spring in the Senate by Robert Bennett (Republican, Utah) is expected to make similar recommendations.

But Representative Jim McDermott (Democrat, Washington), a psychiatrist, plans to introduce a bill that would require explicit informed patient consent for research using identifiable records or samples.

Researchers are very concerned about a bill introduced last month in the Senate by Pete Domenici (Republican, New Mexico), and co-sponsored by James Jeffords (Republican, Vermont). Jeffords chairs the Senate Labor and Human Resources Committee, to which the bill has been referred. The Genetic Confidentiality and Nondiscrimination Act would forbid employers and health insurers from discriminating on the basis of genetic information. It would also require written informed consent for any research use of genetic information, whether or not identifiable samples are involved.

Scientists argue that giving separate protection to genetic information in medical records would be impractical. Some say that the onerous bureaucratic burden the bill would create would paralyse research. But Domenici said that, under his bill, "it will only take 15–20 minutes to inform an individual and get the individual's consent — or not".

Polls show that the American public is concerned about medical records privacy, with 85 per cent calling it very important or essential. Thirty-one per cent say that it would be "not at all" acceptable for medical researchers, keeping patient identities confidential, to use records without consent. Less is known about public attitudes on stored tissue samples. But Gellman said that, on both issues, researchers risk losing a public relations battle. "The research community has a good chance of winning, but they're going to have to work at it," he said. **Meredith Wadman**

Sandra Thurman named as new 'AIDS Czar'

[WASHINGTON] President Bill Clinton this week nominated Sandra Thurman, the former head of a large southern US AIDS organization, as "AIDS Czar" — director of the Office of National AIDS Policy, which coordinates federal government AIDS policy.

Thurman is already a member of the president's Advisory Council on HIV/AIDS, and worked on Clinton's 1992 and 1996 election campaigns.

Thurman was for four years, until 1993, executive director of AIDS Atlanta, the largest southern AIDS service organization. It tripled in size under her directorship.

Some AIDS advocates welcomed her new appointment. "She's a strong choice. She has a good mix of history, political skills, access and tenacity," said David Smith, a

spokesman for the Human Rights Campaign, the largest political gay and lesbian advocacy group.

But another advocate bemoaned what he called the essential powerlessness of the position. "It has no management authority over any program that affects people with AIDS," said Gregg Gonsalves, policy director with Treatment Action Group.

Separately, a draft proposal by Clinton's Advisory Council on HIV/AIDS has called for "a significant increase in funds" for vaccine research and development. "These funds must be derived from new sources within the government and industry," the draft said. It also called for Vice-President Al Gore to head a new public-private vaccine advisory council. **M. W.**

Tobacco industry faces fresh trials as product liability case begins

[WASHINGTON] A seminal trial in the saga of the buffeted US tobacco industry opened in Florida on Monday (7 April), when the family of a woman who died at 49 of lung cancer brought a product liability suit against R. J. Reynolds Tobacco (RJR).

The plaintiff's case relies heavily on internal industry documents which allegedly appear to show that cigarette manufacturers purposely concealed knowledge of tobacco's dangers from the public. RJR has countered that the woman, Jean Connor, had ample warning of tobacco's dangers from numerous sources.

The plaintiff's lawyer has said he will use damaging, newly revealed RJR documents in which a company researcher wrote in 1962 that "obviously" the evidence that cigarettes are damaging to health was "overwhelming". In a 1972 document, another RJR official allegedly wrote that "happily" for the industry, nicotine is "habituating".

Japanese physics centre heads for Brookhaven

[WASHINGTON] The Japanese Institute of Physical and Chemical Research is to establish a new physics centre at Brookhaven National Laboratory on Long Island, New York, under the directorship of T. D. Lee, a professor at Columbia University and a Nobel laureate.

The centre, which will receive \$2 million in Japanese funding in its first year, will concentrate on supporting young physics postdoctoral students and fellows and on exploiting the \$500-million Relativistic Heavy Ion Collider, which comes on stream at Brookhaven in 1999. Japan has provided \$20-million worth of hardware to support the construction of the collider.

Research at the centre will focus on theoretical physics initially, but will move on to experimental work in future years, when funding is expected to increase substantially.

Scandinavia plans park for medical technology

[LONDON] Universities, business and the health sector in Sweden and Denmark are preparing to designate regions either side of the Baltic Sea as a 'medical technology park', designed to promote industries related to biotechnology and medicine.

The 'Medicon Valley' will be linked by a 16-kilometre road and railway bridge joining Copenhagen with Malmö in Sweden, and scheduled to open in 2000. Last week, the backers of the enterprise set up the Medicon Valley Academy with an initial budget of

DKr20 million (US\$2.9 million).

The academy's founders say they hope that the bridge will stimulate cross-border contacts between industry and universities. The proposed area includes two universities — Copenhagen and Lund — three teaching hospitals, and many prominent medical and biotechnology companies, including Astra and Novo Nordisk.

NASA severs links with space probe Pioneer

[LONDON] NASA, the US space agency, has severed links with its spacecraft Pioneer 10, which was launched a quarter of a century ago and placed in a trajectory to escape the Solar System.

At 6.22 billion miles from Earth, Pioneer 10 is the furthest manmade object in space. But it had only one experiment still working, and a weak signal that took the best part of a day to reach Earth.

Since its launch in March 1972, Pioneer 10 took the first close-up photographs of Jupiter and its moons, successfully navigated the asteroid belt and used a planet's gravity to change its course. The spacecraft carries a plaque designed by the late Carl Sagan, which depicts a man, a woman, and a map of the Solar System, designed to show aliens where the spacecraft originated should it fall into the hands of another civilization.

Construction delayed at ignition facility

[WASHINGTON] Construction of the National Ignition Facility at Lawrence Livermore National Laboratory in California, which was due to begin last week, has been postponed until May. A spokesman for the laboratory said that the delay arose to accommodate guests who want to attend the ground-breaking ceremony and would not affect the construction schedule. A group of environmentalists who plan to challenge the construction of the facility in court (see *Nature* 386, 427; 1997) said that they would also delay their planned court action.

Brain science conference comes to White House

[TOKYO] The United States has agreed to host an intergovernmental conference to discuss the future of the Human Frontier Science Program, the international research programme in molecular biology and brain science that was launched in 1987, on a Japanese initiative.

A key item on the agenda at the meeting will be future funding of the program. Japan, which currently contributes 80 per cent of the annual budget, hopes to address this funding inequality at the meeting.

The meeting, which will take place at the

White House on 20 May, will be attended by representatives from the United States, Canada, France, Germany, Italy, Switzerland, the United Kingdom, Japan and the European Commission.

Tuberculosis advance halted, claims WHO

[LONDON] The World Health Organization claims to have halted the upward surge in the global incidence of tuberculosis, for the first time in decades.

The agency says cases of TB have been cut from 6.5 million annually in 1990 to just over 6 million in 1996. The reduction is attributed to the success of a TB-controlling strategy combining multi-drug therapy with a new health management system.

The strategy, DOTS — Directly Observed Treatment Short-course — has been described as "the biggest health breakthrough of the decade" by Hiroshi Nakajima, WHO's director-general. Nakajima predicts that 10 million TB deaths will be prevented over the next decade as DOTS is introduced in more countries.

Cryptography guidelines adopted by OECD

[LONDON] The 29-member Organization for Economic Cooperation and Development (OECD) has adopted a set of guidelines on cryptography policy. Many countries are beginning to formulate policies for data security. But the OECD document, which was published at the end of last month, notes that failure to coordinate these policies will lead to obstacles in the flow of electronic information, and could impede international trade. The guidelines, somewhat controversially, recommend that third parties should be allowed legal access to decryption keys.

Senate urged to ratify chemical weapons pact

[LONDON] US president Bill Clinton has urged the Senate to ratify the United Nations Chemical Weapons Convention, warning that the United States could face diplomatic isolation and trade losses if it fails to do so.

The treaty is set to enter into force on 29 April, as it has been ratified by more than 65 countries. But it faces stiff opposition from conservative Republicans who claim the treaty will be difficult to verify and opens up the United States to intrusive inspections.

If it does not ratify the convention, the United States faces bans on trading in many chemicals and pesticides. In an attempt to address Republican concerns, the Clinton administration has suggested formalizing a policy of retaliating against any chemical warfare attack against the United States.

Time to put malaria control on the global agenda

Growing international awareness of the impact of malaria, and in particular the prospect of an imminent catastrophe in Africa, is generating an unprecedented groundswell for a bold new effort in control.

Exactly 100 years after Ronald Ross discovered the role of the mosquito in the life cycle of the malaria parasite, efforts to control the disease stand at what could prove to be a historic watershed. Over the past year, a number of research organizations, led by the US National Institutes of Health (NIH) and France's Institut Pasteur, have been meeting with malaria researchers, research charities and development agencies, to draw up a coordinated international strategy for malaria research.

The outcome of these discussions is being fed into talks between the World Bank and the African office of the World Health Organization (WHO), which are exploring the feasibility of setting up a multi-agency, 30-year control programme — the African Malaria Initiative. All of these players are in turn talking with the pharmaceutical industry to find ways to reverse its unprecedented withdrawal from vaccine research and development of antimalarial drugs. Testifying to this movement, this week's *Nature* carries an appeal calling on the international community to mobilize in the fight against malaria (see page 541).

Africa's cry for help

The impending crisis in Africa has triggered this resurgence of interest. The parasite's resistance to chloroquine, the main means of prophylaxis and treatment on the continent, is growing alarmingly. Fansidar, the only cheap alternative, seems doomed to the same fate. "It's going to be pretty shocking," predicts Louis Miller, head of the Laboratory of Parasitic Diseases at the US National Institute of Allergy and Infectious Diseases.

A recent seven-fold increase in malaria deaths over five years in parts of Senegal has been linked to emergence of chloroquine resistance, according to Jean-Francois Trape, a Dakar-based researcher from ORSTOM, the

French development research agency. And chloroquine resistance in Senegal is much less intense than in other parts of Africa.

Malaria is already estimated to kill between 1.5 and 2.7 million people every year — several times as many as died in the recent genocide in Rwanda. Another 300 to 500 million people have the disease, and one-third of all humanity lives in zones where they risk catching it. Malaria kills one person — often a child under five — every 12 seconds.

Nine out of ten cases occur in sub-Saharan Africa, and two-thirds of the rest are concentrated in just six countries. These are, in decreasing order of prevalence, India, Brazil, Sri Lanka, Vietnam, Colombia and the Solomon Islands.

Malaria is resurging in many countries

where it had been eliminated or sharply reduced during the 1950s by a WHO-coordinated campaign to eradicate the vector, *Anopheles* mosquitoes, using pesticides such as DDT.

Malaria research and control slipped off the international agenda in the two subsequent decades, says Anatoli Kondrachine, head of malaria control at WHO. The relative success of the campaign lulled many countries into a false sense of security, he says. "Now we are paying a very heavy price for that gap."

Research is helping to fill that gap. New control measures and treatments are most needed in sub-Saharan Africa, where transmission rates are so high that levels of reduction that would bring the disease to a dead stop in India or Brazil have little impact on mortality. The region was in fact left out of the early eradication campaigns precisely for this reason.

Outside Africa, much can be done using existing control techniques given adequate commitment (see overleaf). But research is needed to deploy these measures more effectively, and to develop the new tools needed for comprehensive control. Mosquitoes have grown resistant to pesticides, and the harm that these cause to the environment has restricted their use further. New drugs, and the elusive vaccine, are also needed by all malarious

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

This Briefing has been written by Declan Butler, European correspondent for *Nature*, with additional reporting by John Maurice and Claire O'Brien.

On other pages | India: p536 | Research: p537 | Vaccines: p537 | Vectors: p538 | Money p539 | Industry p540

countries, not to mention the tourists and other travellers who visit them.

Thirty-year offensive proposed

Faced with this situation, Ebrahim Samba, African regional director of WHO and Richard Feachem, director for health, nutrition and population at the World Bank — and former dean of the London School of Hygiene and Tropical Medicine — last year began exploring the idea of creating a multi-agency, 30-year programme to control malaria.

Discussions have since been broadened to include the US Agency for International Development (USAID), US Centers for Disease Control, European Commission, the Organization of African Unity and other agencies. "There should be no delusion that this is a five- or ten-year programme," says Feachem. "We are thinking of a 30-year effort, and everyone would need to buy in for durations of that order."

Over the past decade, countless international and national strategies to "end malaria" have been devised. But many have gathered dust in government drawers, and few have been properly or fully implemented. "The problem is what people are doing in the field, which is not a lot and not very effective," complains Feachem.

The thrust of the new initiative is on building the infrastructure needed for

malaria control in Africa. Back this with a coordinated and adequately funded international effort, and you have the makings of a situation where existing control and therapeutic tools could be used more effectively, says Feachem.

But research would also be an essential component, and the world's major research bodies would therefore need to be "formal partners". Harold Varmus, director of the US



Feachem: backs 30-year effort.

National Institutes of Health, says that joining forces would be "very, very healthy for malaria research in Africa".

The possibility that the programme would fund research directly has not been excluded, says Feachem. "Everything is on the drawing board. The initiative will be an umbrella for many things, and one will need to be a well-directed, adequately funded programme of research; it could be inside or outside, but strongly linked."

None of the agencies discussing the proposed African Malaria Initiative is yet talking dollars. But many of those involved are convinced that the programme will provide a vehicle for better use of existing funds, and act as a magnet to attract new money from the international community.

The initiative has only become possible because of a recent shift in policy at the World Bank. Its support for health projects has leapt from negligible levels in the 1980s to \$2 billion annually — about 10 per cent of the bank's current spending — and is growing. There is said to be consensus within the bank that, among the problems that Africa faces, malaria is near the top.

Political commitment crucial

A decision on whether to launch the African Malaria Initiative is expected later this year, although it would probably not begin until around 2000. "We are in no hurry, we want to get it right," says Feachem.

Political and financial commitment from African leaders is acknowledged to be a prerequisite for the success of the initiative. Its supporters argue that there is no point in external agencies putting money into a sector to which the government of the country is not giving priority.

In practice, health, and malaria in particular, comes near the bottom of the spending priorities of most African countries, and far behind items such as arms, which often eat as much as half of state spending. But change is in the air. When the Organization of African Unity (OAU), meets for its annual summit later this year, malaria will be on the agenda of the 53 heads of state for the first time.

Political commitment is also needed to reverse widespread 'fatalism' about malaria in Africa. "Countries seem to accept that living with malaria is inevitable, that there will always be dead children," says Déogratias Barakamfityé, director of malaria control at the African regional office of WHO. Infant mortality from malaria is so high that in some cultures parents wait a few years before christening their children.

Robert Mshana, executive secretary of the OAU, says that the aim of the summit is to persuade African heads of state to "accept that the [success of the] fight against malaria will be largely dependent on African countries themselves". He adds: "African leaders need to say: 'Malaria is killing our children, we must be determined to do something about it'."

Winning a political commitment would strengthen the hands of health ministers in negotiations with their finance ministries. "We want them to persuade politicians that there are tools which could reduce mortality, that there are considerable economic losses, and that they must allocate the human and financial resources needed," says Barakamfityé.

But Mshana also warns that political will is by itself insufficient. "We have had these slogans for the last 50 years and nothing has moved." Governments could help immediately simply by reducing taxes on products used in research, control or therapy, he says: "In some places the import taxes on mosquito nets are so high that nobody can afford them." □

India plans \$200 million attack on malaria

When India launched a major malaria eradication programme in the early 1950s, 75 million people in the country were catching the disease annually, and 800,000 were dying. Twelve years, and much DDT, later, malaria had gone from 90 per cent of the territory, new cases had dropped to 100,000 annually, and nobody died.

But at the end of the 1970s new cases climbed to 6.5 million annually, and over the past two decades the government has spent up to a quarter of its health budget on malaria control. In the past three years, there have been four major epidemics. Last year 2.85 million cases were reported, and the official — and under-reported — death toll was around 3,000.

Shiv Lal, director of the National Malaria Eradication Programme, accepts that the early successes led to "complacency" in later years. The national malaria control programme was shut down, only to be restarted in 1981 as the NMEP. But today's efforts are being hampered by increasing resistance to chloroquine, the diminishing efficacy and increasing cost of pesticides, and the spread of mosquitoes in urban areas. In addition, *Plasmodium vivax*, which was prevalent in India in the past, has been succeeded by its more deadly relative *P.*

falciparum, which has been less susceptible to chloroquine.

The situation is "bad," admits Vulimiri Ramalingaswami, former chief of the Indian Council of Medical Research, who partly attributes the resurgence of the disease to poorly planned irrigation and urbanization programmes.

India will spend US\$40 million on malaria control this year — up 60 per cent on last year. It is also planning a five-year programme targeting 210 million people in 100 high-risk districts that account for 80 per cent of all *P. falciparum* cases in the country, says Lal. The programme will introduce new tools such as pesticide-impregnated bednets. This initiative will be funded by a loan from the World Bank. The sum has yet to be agreed, but will probably be around \$200 million, says Prabhat Jha, an official at the bank.

Some argue that bureaucracy and not money is the main obstacle to control. Operations are run by the states which have jurisdiction over health matters, and this has relegated the national programme to being a supplier of drugs and pesticides, claims one observer. The national programme "looks good on paper but does not function at the grass roots," he says. **K. S. Jayaraman**

Research 'essential' for control strategy

Two main research goals are most likely to lead in the long term to large strides in malaria control. One is a vaccine, and the second is likely to be the genetic engineering of mosquitoes to confer resistance to the parasite. And there is growing awareness that a more science-based approach is needed in general if malaria control techniques are to be implemented effectively.

One glaring need is better understanding of the complex interaction between malaria transmission rates, acquired immunity, and the clinical forms and incidence of the disease. In regions such as sub-Saharan Africa, where an individual may receive up to 1,000 infectious bites a year, many people slowly acquire a natural immunity that allows them to tolerate a parasite load without developing severe symptoms.

When such populations experience prolonged periods of low transmission, individuals can quickly lose their immunity and risk developing severe malaria if transmission rates increase again. This fact has enormous implications for both the success of control measures, and the understanding of the frequent epidemics triggered by fluctuations in climate.

Some scientists argue that control methods such as insecticide-impregnated bed-nets — which have been shown to cut

mortality and morbidity by 15 to 33 per cent in short-term trials — may in the long term only make populations more vulnerable to severe disease by reducing natural immunity. If they are correct, methods such as bed-nets and genetically engineered mosquitoes would have little impact on the control of the disease, and indeed could even make matters worse.

This complex debate seems to be going the way of those who feel that the lives saved immediately justifies taking the risk. But it has revealed the overriding need to accompany vector control programmes with rigorous long-term epidemiological studies of their impact on immunity, morbidity and mortality.

The same need applies to urban malaria, where transmission seems to be reduced, but severe and fatal disease more frequent. This may have important public health significance, given that migration from rural to urban areas is expected to be the biggest social change in Africa in the coming decades.

The need to understand the complex dynamics of malaria is also acute in other sectors. There is growing awareness that proper testing of candidate vaccines, and successful introduction of any working vaccine, will require much more research on epidemiology, parasite polymorphism, and mechanisms of immunity and transmission,

and knowledge of how these vary from region to region.

More broadly, there is growing awareness that control needs to be more science-based. Richard Feachem of the World Bank says that a comprehensive and integrated approach is needed, not what he describes as "a thoughtless grabbing of the latest attractive idea and just putting it into the field without adequate care".

Harold Varmus, director of the US National Institutes of Health, describes moves to develop control programmes without a strong emphasis on research as "penny wise and pound foolish. I don't think you can do malaria control without a research component, because we don't really know the best way to do control."

Indeed, many believe that what is now needed is an integrated approach, stretching from better diagnosis and treatment of sick children in the home to rational drug use to limit the spread of resistance, monitoring and evaluating control measures, and developing tools such as drugs and vaccines through basic research. The simplistic notion "that a magic bullet developed in a laboratory in London or Washington would bring an end to malaria is now obsolete", says Luiz Pereira da Silva, a vaccine researcher at the Institut Pasteur in Paris.

Vaccines: a roller-coaster of hopes

After many years of frustration, malaria researchers are excited at the prospects raised by a preliminary evaluation of a vaccine developed by SmithKline Beecham Biologicals and the US Army, which protected six out of seven volunteers subjected to repeated bites by infected mosquitoes.

The vaccine is based on the major surface protein of sporozoites, the circumsporozoite protein, which helps them infect hepatocytes. Previous vaccines using this protein had failed in trials because they did not elicit a sufficient immune response. But the new study suggests that this obstacle might be overcome using adjuvants, compounds which stimulate the immune system. "It's a tremendously exciting development", says Ripley Ballou, head of immunology at Walter Reed.

The study tested not just the vaccine but also three different adjuvants. Three groups of subjects were assigned to groups differing only in the adjuvant used. The six out of seven protected were all given a proprietary adjuvant codenamed QS21 (all six control volunteers became infected). This underlines the important role that adjuvants have in protection, not just in vaccines for malaria, but for many other diseases.

Current trials will be watched carefully for signs of progress in the hunt for a vaccine.

The early quest for a vaccine built on the observation that people repeatedly bitten by infectious mosquitoes gradually acquire immunity over a period of many years. Moreover, in the 1960s the British scientists Sydney Cohen and Sir Ian Mc Gregor discovered that antibodies from such individuals could reduce

the parasite load in people lacking immunity and clear disease symptoms.

Around the same time, researchers at New York University (NYU) achieved full protection for the first time, using animal models — by injecting small numbers of sporozoites derived from mosquitoes that had previously been irradiated. Later, researchers at the University of Maryland, NYU and the Walter Reed Army Institute showed that 90 per cent of a group of human volunteers challenged with sporozoites from irradiated mosquitoes later resisted exposure to virulent sporozoites.

Search for 'sporozoite' vaccines

But this was hardly a practical field strategy; inducing resistance required exposing volunteers to several hundred bites by irradiated mosquitoes over several months. The more practical alternative of culturing millions of sporozoites, irradiating them directly, and formulating a vaccine, was also precluded, as sporozoites were impossible to culture (as they still are).

But the successes sent researchers scurrying to their laboratories to find sporozoite proteins that would reproduce the effect of whole

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

irradiated sporozoites. In the 1980s, a group at New York University, showed that monoclonal antibodies against the circumsporozoite protein of rodent and human malaria protected against infection. They later cloned the gene coding for a CS protein, although vaccines based on this failed in trials because they did not elicit high titres of antibody.

Such 'sporozoite' (or pre-erythrocytic) vaccines are just one of three broad vaccine approaches each targeted at one of the three main phases of the parasite's lifecycle. Sporozoites, the form of the parasite injected by the feeding female mosquito, migrate through the bloodstream to the liver within half an hour. In the hepatocytes, they pass through several intermediate stages, and multiply, before emerging around a week later as tens of thousands of 'merozoites' (events up to this point are known as the pre-erythrocytic phase, which is also asymptomatic).

These merozoites invade red blood cells where they complete their cycle, and later rupture the cells and invade new red blood cells (the asexual erythrocyte phase). The 'sexual phase' begins when some of these then form gametocytes — these are sucked up by feeding mosquitos, inside which they combine to form new sporozoites.

Pre-erythrocytic vaccines target the sporozoites themselves or the forms inside liver cells (although the parasites are inside the cell they express epitopes on its surface which can be detected by T-cells that kill the cells directly or by releasing lethal cytokines such as gamma interferon). The liver stages are in principle ideal targets because they remain in the cells for several days. The main difficulty is



that there is as yet no way of making a vaccine which elicits a T-cell immune response.

'Anti-asexual blood stage' vaccines target the phase where merozoites infect blood cells — the phase that causes disease symptoms. One of the most studied vaccines is based on a surface protein of merozoites (MSP1), which gives protection in rodents and primate models — this protein is particularly promising because the regions involved in protection shows no variation. In contrast, another target, EMP-1 (or erythrocyte membrane protein) a parasite protein expressed on the surface of the red blood cell — and which is involved in cerebral malaria — shows high antigenic variation and is not suitable for vaccine development.

Attack on parasite inside mosquito

'Transmission blocking vaccines' interrupt fusion and development of parasite gametes in the mosquito itself. They do not protect individuals directly, but are aimed at reducing transmission, which is why they are sometimes called 'altruistic' vaccines. A fourth approach is therapeutic vaccines targeted against toxins released by the parasites

themselves or during red cell rupture.

Hopes for an anti-asexual blood stage vaccine were raised with a report by Colombian biochemist Manuel Patarroyo of a promising small trial of the controversial synthetic SPf66 vaccine (see *Nature* 332, 158; 1988). Early trials have been described as giving efficacies as high as 80 per cent, but their design has been criticized, and subsequent trials gave efficacies of just 31 and 8 per cent. One last year showed the vaccine had no effect. Patarroyo plans to test his SPf66 vaccine in 50 volunteers using the QS21 adjuvant in Colombia this summer.

In the 1990s, trials of three pre-erythrocytic vaccines in Africa and Asia have also yielded poor results, again because they were poorly immunogenic. A multistage vaccine, NYVAC-Pf7 developed by Virogenetics and Connaught, which showed promise in animal studies, has also given disappointing results. Such multistage vaccines make a concerted attack on several stages of the parasite within the body.

DNA vaccines are another novel approach. Stephen Hoffman, head of malaria research at the US Naval Research Institute in Bethesda, Maryland, plans to test a single-gene vaccine in volunteers this summer. If all goes well, the institute will then test a more sophisticated vaccine devised by Vical, a biotechnology company based in San Diego. The Italian subsidiary of the US company Chiron Vaccines is also working on a multi-stage DNA vaccine.

Apart from the promise of adjuvants, optimism is also high about CS-based vaccines because of the fact that insecticide-impregnated bednets have recently been shown to bring about large reductions in mortality. This suggests that the number of incoming sporozoites is a critical factor in mortality and that even a CS vaccine which did not give complete protection could have a big impact on deaths.

But one caveat in the study by SmithKline Beecham and the Walter Reed Army Institute is that the cut-off point used to define protection was 60 days, and a question remains whether protection will be sustained. A vaccine giving a short-lived effect might protect travellers but not people living in endemic regions. Another question is whether the adjuvants would be toxic in young children. Answers to these and other questions should come from vaccine trials under way in the Gambia.

Transgenic mosquitoes: a new solution?

Many scientists believe the time has come for a serious research effort to develop new techniques of vector control. In particular, rather than focussing on eradicating the mosquito, which is notoriously difficult to control, they are now focusing on creating transgenic insects incapable of transmitting malaria, and finding ways to spread the genes involved in natural resistance throughout wild populations.

Such genes are currently being mapped by genetic linkage studies of malaria-resistant strains of mosquitoes. For example, several such genes have been mapped in a refractory strain of *Anopheles gambiae* by Frank Collins, a vector biologist at the Centers for Disease Control and Prevention and his collaborators Liangbiao Zheng and Fotis Kafatos at the European Molecular Biology Laboratory in Heidelberg.

Collins and Kafatos have collaborated in a multilaboratory effort to map the mosquito genome, which began in the late 1980s with support from the US MacArthur Foundation and Britain's Wellcome Trust.

Another approach is to engineer mosquitoes carrying foreign genes that confer protection against infection. Julian Crampton and his colleagues at the Liverpool School of Tropical Medicine in Britain have cloned a mammalian gene for an antibody that attacks an antigen on the ookinete phase of the parasite in the mosquito.

The bigger challenge perhaps is to drive such engineered traits into the wild population. Work on *Drosophila* has shown that engineered mobile portions of genes — or transposons — can integrate and be replicated in the genome of germline cells and spread quickly through the population. Releasing of a small number of transformed mosquitoes should result in the desired genes spreading throughout the wild population.

No one has transformed mosquitoes in a stable fashion using transposons. But researchers are optimistic that they are close to reaching this goal, while remaining aware that many safety issues need to be resolved before genetically engineered mosquitoes can be safely released into the environment.

Global agencies hold the financial key

The David now challenging the Goliath of malaria is particularly diminutive. Global spending on malaria research by the public and non-profit sector is estimated at just US\$84 million annually¹ and, even after adding industrial spending, the total would probably be no more than twice that. Spending on cancer research in the United Kingdom alone is double that spent of malaria research worldwide.

Furthermore the funding curve over the past decade has been downwards. The US Agency for International Development, the biggest single funder of malaria research, reduced its support from \$49 million in 1985 to \$9.7 million in 1994 (see Figure). Competing demands for funds have spiralled since the end of the Cold War, claims Carter Diggs, a USAID official, in particular support for the newly independent states of the former Soviet Union.

Some of this funding gap has been plugged by increases at Britain's Wellcome Trust and the US National Institute of Allergy and Infectious Diseases. And many are now confident that the nadir has been reached and that an upturn is about to take place, given the renewed interest in the disease; WHO will this year inject an extra \$10 million into malaria research and control, for example, a sum equivalent to one-third of its annual spending in this area.

Scientists' hopes are also being boosted by the prominent lead being taken by Harold Varmus, director of the US National Institutes of Health (NIH), in promoting malaria research. Varmus says he is keen for NIH to "play a bigger role in health in Africa. I find it very easy to make the case that the Departments of State and Defense and development organizations need to think about the rewards of medical research and better healthcare and its impact on promoting stability in Africa."

Greater US involvement in Africa would "really help", says Keith McAdam from the UK Medical Research Council's (MRC) centre in the Gambia, arguing that in the past the United States has not been prominent in setting up long-term commitments on the continent.

Varmus's aspirations are backed by a report released last month by the US Institute of Medicine's Board on International Health, which calls on the United States to take a stronger lead in global health issues by coordinating work among academics, non-governmental organizations, industry and international agencies. It is in "America's enlightened self interest to more fully engage in global health", the report concludes.

The report reveals that US support for international health projects is lower now as a proportion of gross domestic product than at any time since 1950. Its overseas assistance lags behind that of several other industrialized countries. In 1995 the United States spent \$7.3 billion on overseas health projects, compared with Japan's \$14.5 billion, France's \$8.4 billion and Germany's \$7.5 billion.

The spirit of Dakar

The international scientific community is now making a new effort to coordinate malaria research and control and attract new funding. Its efforts crystallized at a meeting, 'Malaria in Africa: Challenges for Cooperation', held in Dakar, Senegal, in January. There, for the first time, the world's research agencies, charities and major donors — such as the World Bank — sat around the same table to explore ways forward. Varmus and Maxime Schwartz, his counterpart at France's Institut Pasteur, were the driving forces behind the meeting.

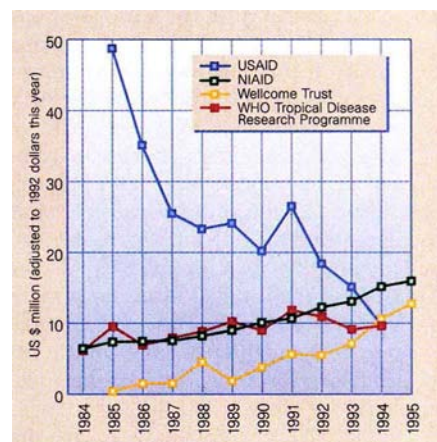
The meeting debated ways to reinforce and coordinate research needed to develop or improve tools for malaria control. One consensus reached was the overriding need to build sustainable research capacity in Africa. This emphasis on African science was reflected in the unprecedented number of African researchers participating.

"If this meeting had been held in 1980 you would have had much fewer scientists from Africa; it would have been driven from the North," says Tom Nchinda, who recently retired as head of research and capacity building in Africa at WHO's Tropical Disease Research programme. "Now we have a very substantial crop of well-trained African scientists."

Research agencies in the North have closed many field stations in developing countries over the past two decades, and the immediate need is to preserve established centres of excellence in Africa, argues McAdam. "It is easy to obtain three-year grants but that doesn't create the infrastructure to make a place work," he says. Stable funding from the MRC has been essential to the survival of the Gambia centre, he says, given that the \$4 million annual grant accounts for almost three-quarters of its budget. But such sustainable support is "difficult to find" at present, he says. The centre has about 500 scientists, 430 of them African.

The Dakar meeting also identified Internet access and other communications infrastructure in Africa as an immediate priority, as this would help scientists to work together more quickly and efficiently, and to share data using common protocols. Varmus pledged an immediate \$1 million and the help of NIH's technical expertise.

"A small amount can go very far here,"



Ups and downs: increased support from research bodies has, until now, not compensated for a fall in USAID funds.

says Varmus. McAdam agrees: "It would be an enormous boost if the US National Library of Medicine could be on-line here. Communications at field stations are poor."

Another small-scale initiative that could have an important impact is the Malaria Foundation — the brainchild of a small group of scientists who aim to coordinate research and to raise funds from the public. The foundation, now being set up, will use the Internet (<http://www.malaria.org>), set up working groups covering all areas of research and compile a global directory of those involved in the fight against malaria.

The organizers of the Dakar meeting are keen that it should not be a one-hit wonder, and have therefore issued a call for concrete ideas on how further progress can be made. These will be reviewed by an international group of working parties at a follow-up meeting in The Netherlands in July. Scientists whose applications are approved will be invited to submit full proposals.

What will happen next remains unclear. The movement behind the Dakar meeting has no name, and some scientists are hesitant to give it an administrative structure. Varmus says one possibility would be simply to "look at proposals and assign them to different agencies for support". But he says that he is also willing to consider a more formal international structure to pool funds.

Varmus and others have their eyes on attracting funding from new and much larger sources than those already available: development agencies such as the World Bank, whose budgets for development are an order of magnitude higher than those for research. Tapping even a small proportion of the enormous sums available for development would be a major boost to the pitiful sums currently available.

1. Malaria research: An audit of international activity. Wellcome Trust Unit for Policy Research in Science and Medicine.



Varmus: seeks role for NIH in Africa.

Can industry be wooed back into the act?

There is a cruel irony in the fact that, just as the explosion of parasite resistance to drugs makes it more pressing than ever to find new drugs and vaccines, the pharmaceutical industry is abandoning malaria research.

Tropical disease research has been a major casualty of the consolidation of the industry over the past decade, with companies tending to concentrate their resources on fewer — and potentially more profitable — areas. Malaria research was shed, for example, in the merger between Rhône-Poulenc Santé and Rorer in 1990, with the new company Rhône-Poulenc Rorer focusing on just six therapeutic areas.

The number of companies carrying out research on malaria vaccines can now be counted on the fingers of one hand. And industry has virtually abandoned drug discovery — the only company still pursuing such research, Hoffmann-La Roche, says that it too has now decided to stop.

Companies are restricting their activity to the clinical development of those compounds that are known to work. "It's cheaper; we start with a molecule that works," says one industrialist. "It's development, not research."

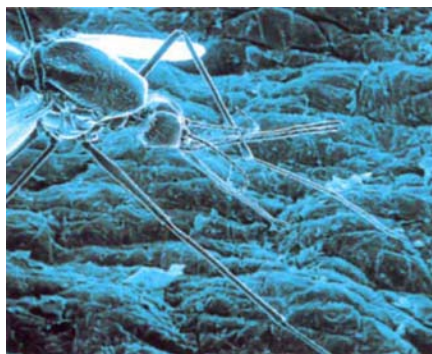
Many companies have had their fingers burned. Over the past few decades, for example, several of them screened hundreds of thousands of compounds without much success. But Pierre Druihle, a researcher at the Institut Pasteur in Paris, argues that it is ironic that companies should abandon drug discovery just when assays for screening for active drug compounds are much more sensitive than in the past, and when genome research and molecular studies are yielding new targets for rational drug design.

A fund has now been created to support a proposed collaborative programme to sequence the 30-megabase genome of *P. falciparum*. The US Department of Defense has given \$15 million to the fund, with another \$4 million coming from the US Burroughs-Wellcome Fund, the UK Wellcome Trust and NIH. Pilot studies are being carried out by the Institute for Genome Research in Rockville, Maryland, Britain's Sanger centre, and NIAID.

It is urgent to relaunch drug discovery, says Dyann Wirth, a molecular parasitologist at Harvard School of Public Health, and chair of WHO's committee on malaria drugs. But without industry, drug development is a lost cause, says Druihle, as only it has the capacity to synthesise and test the hundreds of analogues needed to convert a promising molecule into a drug lead.

Vaccines await 'breakthrough'

The picture for vaccine scientists is less bleak, but only marginally so. The large vaccine manufacturers, SmithKline Beecham and



Once bitten: most companies have abandoned their attempts at drug discovery.

Pasteur Merieux Connaught, still work on malaria vaccines, but many others, such as Merck, Sanofi, and Behringwerke, have given up.

Companies are traditionally reluctant to invest in vaccine development, which carries high costs, low profits and big risks of liability claims. But many companies say that whereas basic research is now producing interesting leads, it has not progressed sufficiently, and that investing now would therefore be a costly and risky business.

The relative lack of industrial interest means that vaccine scientists often cannot produce the large quantities of good manufacturing-practice (GMP) grade antigens needed for clinical trials. "We have six or seven interesting molecules, but have to synthesize them ourselves, while we don't have the competence to work with adjuvants [compounds included in vaccine formulations, which boost the immune system]," says Luiz Pereira da Silva, a vaccine researcher at the Institut Pasteur. "We are forced to improvise, and we can't advance unless we can do trials."

Relaunching industrial research

Faced with the reality of industrial withdrawal from the field, scientists and international organizations are engaged in a flurry of discussions with industrialists to try to encourage them to change their mind.

Research agencies are contemplating strategies to fill the gap left by the retreat of companies, by going beyond their traditional role of funding basic research, and also carrying out development work to bring malaria vaccines to a stage where companies might regain interest.

The idea of creating a broad multi-disease programme to develop vaccine concepts is "under consideration" at the US National Institutes of Health, according to its director Harold Varmus. "We need to try to go as far as possible by ourselves," says Maxime Schwartz, director of the Institut Pasteur. He adds that a "major problem" in doing so in

Europe is the lack of a plant to produce GMP-grade vaccines for research. "This is too expensive for Pasteur alone."

Another idea being floated is for the World Bank and other agencies to subsidize the early stages of drug and vaccine discovery to spread the financial risk. Another is to create consortiums between drug companies, donors and public research agencies. The introduction in Europe of legislation similar to the US Orphan Drugs Tax Credits would also help, says Walter Van Der Smissen, head of government affairs at SmithKline Beecham Biologicals.

Many industrialists say they are open to the idea of taking part in consortia. But they point out that such schemes may not be feasible, given that they are fraught with problems such as sharing of intellectual property, and are likely to fall foul of legislation banning cartels. "I have difficulty seeing how it could happen," says Van Der Smissen, "but perhaps it's time to change the model of thinking."

A more market-driven approach is to guarantee companies sales for their products. There is a precedent. The enormous size of the vaccine market in the developing world — about 3.5 billion doses annually — has allowed organizations such as the United Nations children's fund, UNICEF, to purchase vaccines at cost price. Manufacturers benefit because the huge volume of orders allows them to cut their costs.

Taking the idea further, some suggest that international organizations could specify in advance the characteristics of a specific vaccine or drug which, if met, would qualify it for worldwide use. This would entice companies to invest in the development of such products, say industrialists.

But companies themselves also "need to reassess their markets", argues Richard Feachem of the World Bank. He points out that many countries in Asia and Latin America are relatively well off, while India, and even many poor African countries, have burgeoning middle classes with money to spend on healthcare.

Drug companies are beginning to acknowledge that most of their sales growth over the next ten to twenty years will come from developing countries. "Glaxo-Wellcome and others will put greater emphasis on developing countries and this will lead to renewed interest in meeting local country needs," says one company official.

Growth in these markets will eventually stimulate drug companies to reinvest in tropical disease research. But, until such market forces come into play, companies are unlikely to restart tropical disease research unless the international community provides them with the incentives to do so. □

The spirit of Dakar: a call for action on malaria

Sir — Malaria has been a problem for too long. More than a million people still die from it every year, most of them young children. Nine out of ten of these deaths occur in Africa, where resistance to chloroquine is spreading, diminishing the impact of a drug which in the past has helped to limit the damage wreaked by the disease. Resistance is also increasing to Fansidar, the only other drug affordable for widespread use in Africa. This grim situation prompted an unprecedented meeting earlier this year in Dakar, Senegal, where, for the first time, scientists from both French- and English-speaking parts of Africa and scientists from the North came together to confront malaria.

The challenge is enormous. New approaches are needed to control malaria and its mosquito vectors. New drugs must be found to prevent and treat the disease, and vaccines need to be developed. This requires research at every level; these efforts benefit from cooperation. One example of an international cooperative effort is the sequencing of the *Plasmodium falciparum* genome which will accelerate drug and vaccine discovery and may encourage bright young minds into the field.

The urgent need is to put malaria on the scientific, media and political agenda, and in particular to identify it as a priority for research, both in the developed North and in those areas of the South where the disease is endemic. At the same time, the resources to pursue these goals will remain restricted.

Research is critical for the development of tools for disease control and for their effective deployment. In smallpox, for example, even with the vaccine, the critical insight that it was necessary to vaccinate people only in the vicinity of cases, not the entire population, resulted in cost-effective deployment of staff and resources. In malaria, where the biology of the adversary is more complex, and the existing tools less effective, the interplay between research and control strategies is even more crucial.

The Dakar meeting endorsed the principle that sustained research on malaria in Africa should be a key component of any strategy for tackling the disease. More opportunities for training African scientists as well as for strengthening the research infrastructure will be crucial for building collaborations between African scientists and their partners in the North. Trained African scientists will also be critical for future field trials and interventions.

One unprecedented opportunity to boost

research in Africa at a relatively modest cost is offered by the Internet. An immediate challenge is to identify the systems and resources needed to provide research centres in Africa with access to e-mail and other Internet services. This would allow African scientists to access the scientific information they need, and to communicate quickly and cheaply with colleagues elsewhere.

Electronic networks are an essential aid in building collaborative research networks, for which the need is absolutely compelling if we are to tackle malaria — a disease characterized by highly variable patterns of transmission, pathology and drug resistance in different parts of Africa.

It is only by creating joint repositories of materials — such as strains, reagents and insects — and by common protocols that observations made at various locations and laboratories can be compared. Multi-centre clinical research is also needed to develop or improve therapies for complications of malaria, such as anaemia and cerebral malaria.

Recognizing these needs, the Dakar meeting proposed that sponsoring agencies should promote collaborative research efforts both among African scientists and between scientists in the North and South. An announcement has been made about letters of interest¹ for collaborative research projects. These will be reviewed in July.

A few examples of priorities² for long-term research are provided here.

- Vaccines could be the most cost-effective way to control malaria, but we need both research in the North and collaboration with African scientists to develop candidate vaccines which can be tested in centres in Africa. Studies on epidemiology, parasite polymorphism and mechanisms of immunity in Africa are also a prerequisite for any trials.

- Antimalarial drugs must be developed to replace those becoming obsolete. This will require incentives to attract participation by the pharmaceutical industry and novel approaches to drug development, including production and testing of drugs in Africa.

- The recent development of pyrethrum-impregnated bed nets holds great promise for controlling malaria, but the social and economic factors for their introduction need to be studied in detail. And bed nets cannot be simply introduced, but need to be accompanied by long-term studies to assess their impact on mortality and mosquito resistance to pyrethrum.

- Similarly, case management — recognizing

that a sick child has malaria and providing effective treatment in the home — requires contributions from both the social and biological sciences.

The coming months offer what seems to be a window of opportunity for boosting international cooperation in malaria research as a component of overall control. The African regional office of the World Health Organization with support from the World Bank is considering the conclusions reached at Dakar as part of the development of a long-term, African-developed and African-led malaria control strategy. The Organization of African Unity (OAU) will hold its annual meeting in June, and malaria will be on the agenda. The same month, the G-7 countries will discuss emerging and re-emerging infectious diseases, including malaria.

In the spirit of the Dakar meeting, we recognize that control of malaria in Africa will require a long-term collaboration between scientists in the North and South and the involvement of many more countries than the initial group which took part in the Dakar meeting. It will require commitments from the industrialized countries to funding, and from African leaders to support scientists and health and research infrastructure in their countries.

The fact that the Dakar meeting was opened by the President of Senegal and closed by its Minister of Health seems to symbolize the growing awareness by leaders in Africa of the need to support research efforts by both indigenous scientists and their colleagues abroad. The international community has a responsibility to recognize malaria, like the AIDS pandemic, as a major challenge and to commit itself to do more to prevent the disaster that looms over Africa.

Jean-Marie Bruno (French Ministry of Cooperation); **Richard Feachem** (World Bank); **Tore Godal** (TDR, WHO); **Thomas Nchinda** (elected representative of African scientists for post-Dakar efforts); **Bridget Ogilvie** (Wellcome Trust); **Barend Mons** (Netherlands Organization for Scientific Research); **Robert Mshana** (Organization of African Unity); **George Radda** (Medical Research Council, UK); **Ebrahim Samba** (WHO, Brazzaville); **Maxime Schwartz** (Institut Pasteur); **Harold Varmus** (National Institutes of Health, USA); **The Dakar Scientific Advisory Committee: Samba Diallo** (Senegal), **Ogobara Doumbo** (Mali), **Brian Greenwood** (United Kingdom), **Wenceslaus Kilama** (Tanzania), **Louis H. Miller** (USA), **Luiz Pereira da Silva** (France)

* Correspondence for the group should be addressed to: Louis H. Miller, Bld. 4, Room 126, NIAID, NIH, Bethesda, Maryland 20892, USA; fax (301) 402-0079; e-mail: Louis_Miller@nih.gov

1. Request for letters of interest:

http://www.niaid.nih.gov/dmid/mal_itr_en.htm

2. Priorities from the Dakar meeting can be obtained at:

<http://www.niaid.nih.gov/dmid/malaftr/>

Value for money

Sir— Scientists are currently judged by the number and quality of their publications¹. Using such criteria, Robert May's "Review of world science"² finds that the economically developed nations fare well, whether the results are expressed in terms of citations per person or a relative citation index. In today's economic climate, he suggests that the United Kingdom obtains the "best value for money" at 168.2 citations per million pounds sterling. Such a figure implies that each publication costs £5,945 or US\$9,530. May further suggests that better value for money is associated with university institutions.

Several of my South African colleagues in the life sciences at university institutions have provided me with their average costs per publication. These range from \$500 to \$2,400. The average is \$2,000. This is not a very realistic figure because chemicals are purchased from outside South Africa and are subject to 10% import duty and 14% value-added tax. So the real cost for each publication is \$1,520.

Sceptical of the quality of these publications? These colleagues are publishing in journals with impact factors from 1.7 to 5.2. This suggests that we are

producing 6 papers for each produced in the United Kingdom, and 7 for each produced in the United States. If we include the cost of the manpower involved, we arrive at 24 and 28 publications respectively for the same investment — South African lecturers receive \$14,000 a year (less than a quarter of salaries on offer in the United States as judged by classified advertisements). I should like to suggest that funding agencies in Europe and the United States should invest in South African science in our universities for real "value for money".

J. P. Dean Goldring

*Department of Biochemistry,
University of the Witwatersrand,
1 Jan Smuts Avenue,
Johannesburg, South Africa
e-mail: 089goldr@cosmos.wits.ac.za*

1. Lewison, G., Anderson, J. & Jack, J. *Nature* 377, 671 (1995).

2. May, R. *Science* 275, 793–796 (1997).

History of parity violation experiment

Sir— The Commentary article¹ "Parity and chivalry in nuclear physics" refers to the two first publications demonstrating the non-conservation of parity in the weak interactions^{2,3}. The article is summarized by

the subhead "Forty years ago, the world of physics was stunned by the discovery that nuclear beta-decay does not respect symmetry between left and right. But the credit for this conclusion has not been properly attributed." The purpose of the Commentary was "... to state for the record that the NBS [National Bureau of Standards] parity violation experiment was a collaborative team effort in which nuclear physicists and cryophysicists pooled their knowledge and expertise to carry out an experiment proposed by Lee and Yang, thus confirming their hypothesis that parity is not conserved in β -decay".

We have always regarded this epochal experiment as a team effort. When we wrote in 1957, "[w]e are also indebted to Professor C. S. Wu for reports of her preliminary results in the Co-60 experiment which played a crucial part in the Columbia discussions immediately preceding this experiment", we did not intend to apportion credit among the authors of the Letter reporting the Co-60 results. Our own experiment, "Observations of the failure of conservation of parity and charge conjugation in meson decays: the magnetic moment of the free muon", was spurred by Wu's Friday-lunch report of the status of the Co-60 experiment and was performed that Friday night, 4 January 1957, to



So far, systems in the ÄKTA design family include:

- ÄKTAexplorer, for method development and scale up of every biomolecule
- the new ÄKTApurifier, for purification of peptides, oligonucleotides and other biomolecules

Tuesday morning 8 January, and our Letter was written that day. Had we heard the word from Ambler, we would have thanked "...Dr E. Ambler for reports of his preliminary results..." and we could as well have written the phrase without the "her". As for the 1978 Wolf Prize to Wu, citing the Co-60 experiment as "her most famous work", a prize for lifetime achievement awarded to Ambler could properly cite the same experiment as "his most famous work".

Wu died in New York on 16 February 1997. At a memorial service on 22 February, every speaker gave credit for the Co-60 experiment both to Wu and to the NBS team. For instance, C. N. Yang said: "In 1956 she and her collaborators at the NBS did one of the most exciting measurements..."; and T.D. Lee: "Forty years ago she and her colleagues at NBS overthrew the principle of conservation of parity."

In 1973, Wu herself published a marvellously informative and warm discussion of the Co-60 experiment⁴, followed by a briefer presentation by one of us of the meson experiment⁵, and also by V. L. Telegdi⁶.

In that account, Wu details the origin of the Co-60 parity experiment in early spring 1956 when T. D. Lee described to her the possibility that the τ - θ particle decay anomaly could be due to the violation of

parity conservation in the weak interactions. Wu suggested to Lee that the best bet for testing this hypothesis in β -decay was demagnetization-cooled Co-60, and she decided to work full-time on that. In preparing for the experiment, she and her Columbia colleagues immediately remeasured the spin of Co-60 and, on 4 June, Wu called Ambler to find out whether he would be interested in a collaboration; Ambler accepted with enthusiasm, and work began at NBS. Before going to NBS for the first time in mid-September, Wu had done experiments at 4 K on the detector designed to detect β particles from a source at milli-Kelvin temperature, had made detailed studies of magnetic field effects on the β counting and had studied back-scattering from the cerium magnesium nitrate (CMN) crystal source backing.

Wu also grew two Co-60 specimens, one with the β -emitting thin surface layer containing a few microcuries of Co-60, the other with Co-60 throughout the crystal for preliminary studies of γ anisotropy from polarized nuclei. But the surface warmed rapidly because of condensation of residual helium, and the NBS team used a CMN shell to shield the surface. By Christmas Eve, the electron asymmetry was "reproducible and huge" but rigorous checks were still to come. (R. P. Hudson says that according to the log books, this

occurred on the night of 26 December, personal communication 13 March 1997.) Wu writes: "The period between January 2nd and January 8th was probably the most tense in our whole experimental venture", but these meticulous experimenters finished all the experimental checks and gathered at about 2 a.m. on 9 January to celebrate the great event.

Kurti and Sutton write: "In those days it was usual to list authors in alphabetical order, unless one was the leader of the team or the originator." It would not be amiss to regard Wu as the originator of the experiment, given the facts as related above. But the NBS team of Ambler, Hayward, Hoppes and Hudson, as well as Wu, were full collaborators and deserve full credit.

Richard L. Garwin

Columbia University,
New York, New York 10027, USA

Leon M. Lederman

Illinois Institute of Technology,
Chicago, Illinois 60616, USA

1. Kurti, N. & Sutton, C. *Nature* 385, 575 (1997).
2. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. & Hudson, R. P. *Phys. Rev.* 105, 1413 (1957).
3. Garwin, R. L., Lederman, L. M. & Weinrich, M. *Phys. Rev.* 105, 1415 (1957).
4. Wu, C. S. *Adventures in Experimental Physics*, Gamma Volume 101-123 (1973), from which the information of the next two paragraphs is drawn.
5. Garwin, R. L. *Adventures in Experimental Physics*, Gamma Volume 124-130 (1973).
6. Telegdi, V. L. *Adventures in Experimental Physics*, Gamma Volume 131-136 (1973).

ÄKTAdesign: an open purification platform for all of your biomolecules

What type of purification is going on in your lab? Do some of your colleagues develop methods and optimize schemes to purify peptides, proteins, or oligonucleotides at every purification scale? Are others purifying natural, synthetic and recombinant peptides? Are yet others purifying native or recombinant proteins? Or perhaps you do all of this yourself.

Doing individual types of purification has meant following individual working procedures—until now, that is. Until ÄKTA™ design (ÄKTA is the Swedish word for real; it's pronounced eckta).

**With ÄKTA design, your purification systems
won't act like strangers to one another**

ÄKTA design is the name of a new platform for a family of purification systems and pre-packed columns exclusively from us, Pharmacia Biotech. The platform integrates fully-biocompatible hardware solutions with a control system that gives you control over purification systems from lab to production scales. It lets everyone use the same better, smarter way of doing purification. All of which means you can operate every ÄKTA design system once you've used any one of them.

Each ÄKTA design system lets you use pre-set protocols that automatically resolve all major purification tasks—including automatic method scouting. Each system gives you pre-set running parameters for most purification techniques. Each system is supported with an extensive range of technique-specific, pre-packed columns. Each system automatically prepares buffers from stock solutions—without manual titration. And each system operates via UNICORN®—with this single control system, you can instantly transfer your methods to purification systems at all scales.

What does your lab want to purify today? A version of ÄKTA design will suit all your needs. Call us: 1 (800) 526 3593 from the USA; +81 (0)3 3492 6949 from Japan; or +46 (0)18 16 50 11 from Europe and the rest of the world. Ask for a free brochure. Or meet us on the Internet at <http://www.biotech.pharmacia.se>.



Uppsala, Sweden. (And the rest of the world)

Science as a cultural construct

Scientific knowledge is a communal belief system with a dubious grip on reality, according to a widely quoted school of sociologists. But they ignore crucial evidence that contradicts this allegation.

**Kurt Gottfried and
Kenneth G. Wilson**

The thesis that scientific knowledge is a cultural construct — that, according to Andrew Pickering¹, “there is no obligation upon anyone framing a view of the world to take account of what 20th century science has to say” — is gaining currency. The reaction of scientists to this view of their work received little attention until news of Alan Sokal’s notorious hoax article in the journal *Social Text* hit the front page of *The New York Times* on 18 May last year. But Sokal’s victims are a fringe group, ill-informed about science. We wish to respond to a more sophisticated and troubling view of scientific knowledge as put forward by what we will, for brevity, call the ‘Edinburgh’ school of sociology.

The ‘strong-programme’ Edinburgh sociologists contend that the knowledge produced by the natural sciences is a cultural construct essentially equivalent in its attributes to the knowledge produced by less tractable pursuits in which reproducible phenomena under carefully controlled conditions either do not exist or are of little interest. In making this case, its proponents dismiss or ignore a large body of concrete evidence that contradicts this contention. No view of the world is sound if it ignores the steadily improving predictions of twentieth-century science.

An aphorism of Einstein’s is a good introduction to this topic: “the scientist... must appear to the systematic epistemologist as an unscrupulous opportunist”. Anyone fortunate enough to have been involved in the chaotic outbreak of a scientific revolution will agree. The statement has turned out to have a deeper meaning, however. Scientists eventually settle on one theory on the basis of imperfect data, whereas logicians have shown that a finite body of data cannot uniquely determine a single theory. Among scientists this rarely causes insomnia, but it has tormented many a philosopher (see, for example, ref. 2). In ‘cultural construction’, the Edinburgh school claims to have found the missing epistemological link.

The Edinburgh school does not see itself as opposing science, or questioning the integrity of scientists. But it contends that scientific knowledge is only a communal belief system with a dubious grip on reality, and it is this claim that we address from our perspective as physicists.

Einstein’s aphorism is also the opening sentence of Pickering’s book *Constructing Quarks*, which analyses the birth of the Stan-

dard Model of particle physics. Published over a decade ago, it remains the most instructive, ambitious and outrageous Edinburgh study of modern physics. Pickering displays a solid command of the developments that led to the Standard Model. Naturally, he focuses on cultural factors of which scientists are often unaware, and provides many valid insights into their significance. But consider the conclusion Pickering drew: “The quark–gauge theory picture of elementary particles should be seen as a culturally specific product... a communally congenial representation of reality. [G]iven [its] cultural resources, only singular incompetence could have prevented the high energy physics community producing an understandable version of reality at any point in [its] history... [T]he preponderance of mathematics in particle physicists’ account of reality is no more hard to understand than the fondness of ethnic groups for their native language.”

Methodology applied too widely

How did he reach this remarkable conclusion? A clue is provided by the methodological tenets of the Edinburgh school’s strong programme and, in particular, the requirement of impartiality “with respect to truth and falsity, rationality or irrationality, success or failure”³. Another Edinburgh rule is not to write “Whig history” — historians’ jargon for accounts that allow subsequently accepted knowledge to distort the story, or that assume that progress is inevitable.

At this point, we wish to note an important distinction: ‘science-as-practice’ as compared to ‘science-as-knowledge’, to use Pickering’s terms. The border between these topics is neither sharp nor stationary, but there are times when a clear separation is essential. The strong programme defines a sound methodology for the sociology of scientific practice, for it should not prejudice the case against endeavours that ultimately fail but whose validity and significance are still unknown or ambiguous at the time.

Pickering and others of the Edinburgh persuasion have, with this methodology, produced useful studies of science-as-practice (for example, ref. 4) and are helping to fill an important gap because historians of modern physics have, until recently^{5,6}, paid little attention to experiment while exploring theory in great detail. But Pickering’s conclusions quoted above, and similar statements by others in the Edinburgh mould, emerge because the strong programme’s methodology is applied by them to scientific knowledge itself, and not just to practice.

To turn to the Standard Model, its acceptance in the mid-1970s was a revolution in the sense of Thomas Kuhn’s 1962 classic *Structure of Scientific Revolutions*. The ‘old’ physics was a set of loosely connected and partially ambiguous paradigms: quantum electrodynamics (QED), the firmly established theory of the electromagnetic interaction; the ‘S-matrix’ theory of collisions between strongly interacting particles (called hadrons, such as the proton); disconnected therefrom, an ill-defined quark model of the hadron spectrum that violated the principle of relativity; and a manifestly inconsistent theory of the weak interaction with vague hints of kinship to electrodynamics. It was not known how to define this ‘old’ physics unambiguously by a set of consistent equations involving a finite set of parameters accessible to clear-cut measurement.

The ‘new’ physics arose from various sources: experiments that showed that the proton has a granular substructure; the invention of generalizations of Maxwell’s electromagnetic theory that offered the first mathematically sound formulations of the weak interaction and incorporated QED; experiments that revealed a new type of weak-interaction phenomenon predicted by the simplest of these electroweak unification schemes; the discovery of the new family of ‘charmed’ hadrons with the properties called for by that very scheme; and the proposal of quantum chromodynamics (QCD), according to which hadrons are composites of quarks interacting through a field which is also governed by a similar generalization of Maxwell’s equations. Initially, only QCD had no direct empirical evidence in its favour.

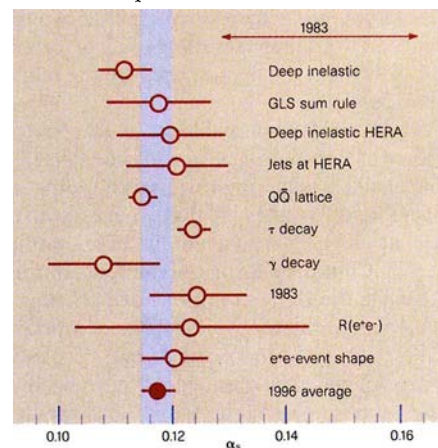


Figure 1 The coupling constant α_s of QCD, the analogue of the fine structure constant ($1/137$) of QED, as found today from a variety of experiments and lattice calculations²¹; and the range of values as determined in 1983.

In contrast to the old physics, the new is a clear-cut paradigm — a firm set of equations that conforms to the requirements of relativity and contains a strictly limited set of unambiguous parameters. Unfortunately, this quite elegant structure, which incorporates not only QED but also atomic and nuclear physics, has been named a ‘model’ instead of a ‘theory’.

That the new paradigm was far more constrained and comprehensive was clear from the start: it was widely seen as a congenial representation of reality, whereas the old one was not. Of greater importance, a stream of new experimental results improved the precision of the evidence in support of the new theory, and confirmed several crucial predictions. The most striking came in 1983, before *Constructing Quarks* went to press, but is barely mentioned in that book: the discovery of the W and Z particles — according to the theory, the mediators of the weak interaction and cousins of the massless photon, yet predicted to have very specific and gargantuan masses, as confirmed by experiment.

The transition from the ‘old’ to the ‘new’ physics is covered in *Constructing Quarks*, much of it very well indeed, but nonetheless with distortions on which the preposterous conclusions hinge. The most important is Pickering’s contention that the shift of focus from the copious phenomena amenable to S-matrix theory to the rare processes of the ‘new’ experimental physics was a cultural bandwagon, because the ‘old’ physics was already a “congenial representation of reality”. It is true that bandwagons tend to plague high-energy physics because so many work on so small a set of phenomena, but this does not prove that in this case people jumped to the wrong conclusion. The S-matrix *Anschaung* had its own bandwagon for a time, but it could not say anything about the proton’s substructure or the weak interaction of hadrons, electrons and neutrinos. There was little ground for hope that another decade of ‘old’ physics experiments would have led to an understanding of these topics, all of which are encompassed by the Standard Model.

Pickering’s statement about the irrelevance of twentieth-century science stems in part from a recurring misunderstanding in sociological studies — that what scientists see as progress often entails abandoning familiar but perplexing phenomena, in this instance the more commonplace phenomena that dominated the ‘old’ physics supposedly abandoned in favour of more esoteric phenomena in the stampede to the ‘new’. But they were not abandoned, they were set aside, as has often been fruitful in physics.

Galileo set aside friction. Bohr’s demonstration that the hydrogen spectrum holds the key to atomic physics led to a shift of interest towards spectroscopy, but after the development of quantum mechanics the

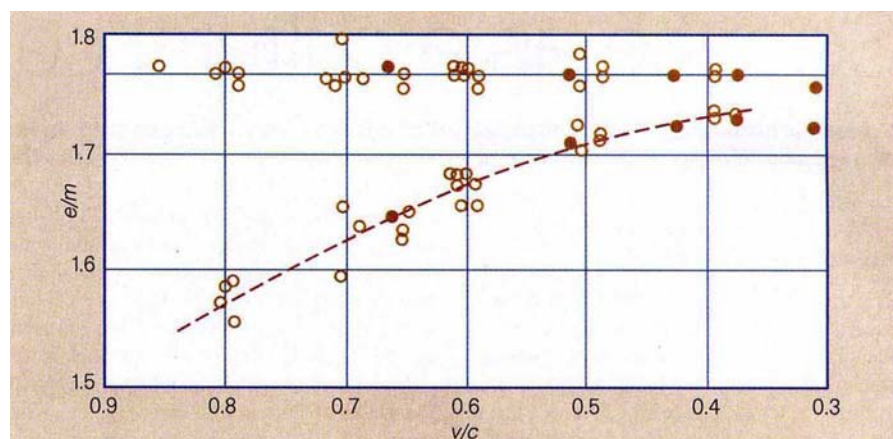


Figure 2 The ratio of the electron’s charge to its rest mass, e/m , as extracted from measurements on radium β rays by A. H. Bucherer in 1909 (filled symbols), and G. Neumann & C. Schaeffer in 1914–16 (open symbols). The vertical scale is $(e/m) \times 10^7$, the horizontal the electron velocity divided by the speed of light²². When the data are analysed with the relativistic formula, they give a constant value (the green line), in excellent agreement with today’s far more accurate value. When analysed with a competing but forgotten theory, the data give the nonsensical (red) curve, whereas newtonian mechanics would produce an even larger discrepancy.

‘abandoned’ phenomena (for example, in solids) returned to centre stage. The ‘new’ particle physics was, in large part, born of the recognition that the ‘old’ phenomenology was less accessible than that on which the ‘new’ physics concentrated. This venerable strategy is being vindicated now with lattice QCD, a numerical approach that is yielding the first fundamental understanding of hadronic spectra⁷ (Fig. 1).

Science as myth

Other insights into the sociology of scientific knowledge can be found in *The Golem: what Everyone should know about science*⁸. We like the title of this book, which was taken from Yiddish mythology, because science is indeed something of an uncontrollable monster created by man. On the other hand, “everyone” cannot assess the book’s validity except by resort to common sense. That may suffice, for the authors conclude that “we have shown that scientists at the research front cannot settle their disagreements through better experiments, more knowledge, more advanced theories, or clearer thinking... [but transmute] wildly varying results... into neat and tidy scientific myth. There is nothing wrong with this. The only sin is not knowing it is always thus.”

The authors cite several case histories in support of this thesis, and here we take those on relativity as examples of this type of argument. The account in *The Golem* of empirical tests of special relativity is based entirely on the Michelson–Morley experiment, because Einstein’s special theory of 1905 is based on the axiom that the speed of light does not depend on the motion of the observer relative to the source. Long before 1905, Michelson and Morley found no effect on the velocity of light from the Earth’s motion through the hypothetical ether.

In 1925, however, and again in 1933, Dayton Miller announced a positive observation, which contradicts relativity, but was ignored. The authors of *The Golem* see this as proof that what matters is not the quality of an experiment but “what people are ready to believe” because “the culture of life in the physics community meant that Miller’s results were irrelevant”. When the word ‘culture’ is replaced by ‘facts’, this sentence is no longer misleading, for in the decade following 1909 half a dozen independent laboratory experiments had confirmed the relativistic relation between velocity and momentum to better than 1 per cent (Fig. 2); in 1923–24 Compton, Bothe and Geiger’s experiments confirmed the relativistic conservation laws for momentum and energy; in 1926 another type of ether drift experiment designed expressly to check Miller’s claim gave a null result; and in 1933 pair-creation, the most extreme manifestation possible of $E = mc^2$, was discovered. Not one of these facts is mentioned in *The Golem*.

A delectable tale about the culture of science, well recounted in *The Golem*, is offered by the ambiguous early tests of general relativity, Einstein’s 1915 theory of gravitation. But then the authors stop the clock at mid-century, leaving the reader with the clear impression that both special and general relativity stand on an empirical house of cards. With this as backdrop, the account closes as follows: “We have no reason to think relativity is anything but the truth... but it is a truth which came into being as a result of decisions about how we should live our scientific lives, and how we should license our scientific observations; it was a truth brought about by an agreement to agree about new things.” What ‘truth’ means here is mysterious. The facts are clear, however, but they cannot be found in *The Golem*: since 1960, a series of

observations using technologies undreamt of in 1915 have confirmed Einstein's three original predictions and produced new precision tests verifying general relativity¹⁰.

The muddle that can result when practice and knowledge are not distinguished is illustrated by the response of Collins and Pinch¹¹, the authors of *The Golem*, to David Mermin's critique of their treatment of relativity¹². Collins and Pinch write that "in exploring science as a craft, [our] strange sociological approach — which temporarily sets aside what we all know to be true — is useful, if uncomfortable", where by 'craft' they refer explicitly to science in the public arena: courtrooms, Chernobyl, and so on. But at issue here is science-as-knowledge in the abstruse case of relativity, not DNA evidence in the O. J. Simpson trial. Moreover, if relativity is known to be true by all, what precisely is it that this piece of knowledge fails to say about Nature?

The interpretation of science exemplified by *Constructing Quarks* and *The Golem* has an interesting relation to neighbouring pursuits. Edinburgh is one rather new and small school within the sociology of science¹³, which is but a small field within sociology. We do not mean this statement pejoratively: the inventors of quantum mechanics also formed a small school. Furthermore, sociologists who study scientific knowledge do not all adhere to the strong programme¹⁴. In Kuhn's terminology, the sociology of scientific knowledge as a whole is in a pre-paradigm phase, like mechanics before Newton when there was no consensus on methods, concepts or goals. Again, this is no insult; it took more talent to be a physicist before the *Principia* than afterwards. Finally, whatever quarrel scientists may have with Edinburgh pales in comparison with the hostilities between Edinburgh and some philosophers of science, a conflict that can make a scientist feel like an Algonquin whose hunting grounds are being fought over by two colonial powers¹⁵.

Another oversimplification

The pragmatic physicist, however, finds that the Edinburgh sociologists claiming to be the bona fide interpreters of scientific knowledge have objectives resembling those of philosophers such as Kuhn. That sociological factors have been important in the development of scientific knowledge was demonstrated by Kuhn, who is recognized as a forebear by the Edinburgh school¹⁶. Both see claims for scientific advances as largely circular arguments because they do not conform to the mythical format in which observation always precedes theory, with logic and rock-solid data always pointing to one, and only one, theory. This impels them to replace 'always' by 'never' in the last sentence, which is yet another oversimplification because the correct word is 'sometimes'. Indeed, these sociologists and philosophers thus have a

relationship to scientists resembling the one between pure mathematicians and physicists, in that most of the latter are satisfied with a level of rigour unacceptable in pure mathematics. Had that not been so, physics would still be struggling with problems solved generations ago.

In terms of this analogy, the Edinburgh position is that of a fictitious mathematician who asserts that the knowledge claimed by physicists is a cultural construct because physicists have a communal agreement to accept standards of mathematical deduction that are patently unsound. Real mathematicians don't make such assertions; they understand that physicists conform to a different but appropriate norm in their practice.

Sociologists as Judge

Many of the misconceptions in the literature on the sociology of scientific knowledge follow from the strong programme's methodology when applied to science-as-knowledge. The requirement that rationality and irrationality, and the like, be treated symmetrically gives equal credence to contending positions of objectively unequal merit — the 'old' versus the 'new' physics in *Constructing Quarks*, and whether cold fusion does or does not exist in *The Golem*. This leaves only psychoanalysis or social construction as explanations for the scientists' choice. Moreover, the sociologist thinking as a historian (who must not write Whig history), but really acting as the judge of scientific knowledge, assumes the power to stop the clock at an arbitrary point, thereby ignoring subsequent evidence as to whether some bandwagon fell over the cliff or stayed on track.

Those who cannot abide the Whig label should issue periodic updates, as is common in the sciences. Our 1986 paperback edition of *Constructing Quarks* has no such addendum, even though the empirical support for the Standard Model had strengthened substantially since the first printing. The crucial parameter of the electroweak theory, which must yield the same value when extracted from a host of distinct processes, was found to be 0.25 ± 0.05 in 1982, followed by 0.223 ± 0.004 (1985) and 0.2236 ± 0.0008 (1996). Indeed, the model's resounding success has kindled boredom in many who entered physics for high adventure. Even QCD, whose uncritical and prompt acceptance was justifiably criticized by Pickering, is prospering (Fig. 1). Proponents of the Edinburgh school would probably say that this only goes to show how powerfully culture shapes what experimenters measure and theorists calculate, which is half-right — but, as with other aspects of their position, half-right is wrong.

Predictive power, the strongest evidence that the natural sciences have an objective grip on reality, is largely ignored by these commentators. For the question of whether scientific knowledge is contingent on culture, the dis-

covery of phenomena that could not have been foreseen when a theory was invented but which are in accord with that theory are especially germane. There are many examples, of which we choose two here. First, two centuries after the *Principia*, Poincaré discovered that Newton's equations of motion describe not only the clockwork Universe that had dominated everyone's world view, but also chaotic regimes. A century later, chaotic motion was identified definitively in the Solar System and found to conform to Newton's original equations¹⁷. Second, almost a century after Planck proposed the black-body law that gave birth to quantum theory, that law was found to fit the cosmic background radiation to better than 0.1 per cent. How can such facts be reconciled with the statement that "the science-as-knowledge image of science is so thin that it bears almost no relation to its subject matter"¹⁸?

Intriguing questions are raised by the complexity of scientific practice, and the Edinburgh school is well-positioned to study them¹⁹. But forcing scientific knowledge into the strong programme's doctrinal straitjacket has often produced sophistry, and compromised studies of practice. The Edinburgh school will only arrive at a more insightful perspective on scientific knowledge if it revisits its underlying assumptions, a difficult task that Kuhn did not shy away from when he found that his reflections had failed to reach convincing conclusions²⁰. □

Kurt Gottfried is in the Laboratory of Nuclear Studies, Cornell University, Ithaca, New York 14853, USA; Kenneth G. Wilson is in the Smith Laboratory, Ohio State University, Columbus, Ohio 43210, USA.

Acknowledgements. We thank Peter Galison, Sorel Gottfried, Henry Kendall, David Mermin, Sam Schweber, George Smith and Scott Tremaine for advice and comments.

- Pickering, A. *Constructing Quarks: A Sociological History of Particle Physics* (Univ. of Chicago Press, 1984).
- Norton, J. D. *Synthese* **99**, 3 (1994).
- Bloor, D. *Knowledge and Social Imagery*, 2nd edn (Univ. of Chicago Press, 1991).
- Pinch, T. *Confronting Nature: The Sociology of Solar-Neutrino Detection* (Reidel, Dordrecht, 1986).
- Franklin, A. *The Neglect of Experiment* (Cambridge Univ. Press, 1986).
- Galison, P. *How Experiments End* (Univ. of Chicago Press, 1987).
- Davies, C. T. H. et al. *Phys. Lett. B* **345**, 42 (1995).
- Collins, H. & Pinch, T. *The Golem* (Cambridge Univ. Press, 1993).
- Chase, C. T. *Phys. Rev.* **28**, 378 (1926).
- Will, C. *Science* **250**, 770 (1990).
- Collins, H. & Pinch, T. *Physics Today* p. 94. (January 1997).
- Mermin, N. D. *Physics Today* March, p. 11. (1996).
- Zuckerman, H. in *Handbook of Sociology* (ed. Smelser, N. I.) 511–574 (Sage, Newbury Park, California, 1988).
- Shapin, S. *Annu. Rev. Sociol.* **21**, 289–321 (1995).
- Symposium on Deconstructing Quarks *Soc. Stud. Sci.* **20**, No. 4 (1990).
- Barnes, B., Bloor, D. & Henry, J. *Scientific Knowledge: A Sociological Analysis* (Univ. of Chicago Press, 1996).
- Wisdom, J. *Icarus* **63**, 272 (1985).
- Pickering, A. *Soc. Stud. Sci.* **20**, 685 (1990).
- Pickering, A. *The Mangle of Practice* (Univ. of Chicago Press, 1995).
- Kuhn, T. S. *The Trouble with the Historical Philosophy of Science* (Dept of the History of Science, Harvard Univ., 1992).
- Phys. Rev. D* **54**, No. 1, p. 80 (1996).
- Gerlach, W. *Hdbch. d. Physik* **XXI/1** p. 80 (1933).

After slip no longer an afterthought

Charles DeMets

Measuring the imperceptible motion of tectonic plates has become possible with sub-centimetre positioning from satellite data. Such measurements show that postseismic creep can release as much energy as violent earthquakes.

If you have ever felt the eerie ground-shaking that accompanies an earthquake, then you are one of millions of people who have briefly experienced the dance of the tectonic plates that slowly drift across the Earth's surface. But earthquakes account for only a fraction of the plate displacements that are predicted by models of present-day plate velocities, thereby indicating the existence of a 'seismic slip deficit'^{1,2}. Using a relatively new technique for measuring active crustal deformation, Heki and colleagues (page 595 of this issue³) demonstrate that seismically invisible slip following earthquakes accounts for some of this deficit. Their work, when combined with other observations of strain release along plate boundaries, bolsters the case that steady motion between plates is accommodated by a continuum of processes that include slow-rupture earthquakes⁴⁻⁶, silent earthquakes⁷, aseismic creep⁸ and afterslip^{9,10} (Fig. 1).

During a typical earthquake, crustal strain that has accumulated over decades or longer is released in a period of seconds to minutes. Seismic waves radiating out from the earthquake source region are recorded by seismometers, instruments designed to measure ground motions and accelerations induced by high-speed fault ruptures. Seismic data collected over the past century have yielded a plethora of information about the earthquake source process, the structure of the Earth's interior and the nature of present-day crustal deformation.

In contrast, attempts to study slip processes that occur over periods longer than a few minutes have been hindered by the lack of an inexpensive and reliable technique for measuring long-term ground displacements. The completion of the Global Positioning System (GPS) satellite constellation in the 1990s dramatically improved the prospects for such studies by enabling sub-centimetre determinations of the absolute and relative locations of points nearly anywhere on Earth. Since October 1994, the Geographical Survey Institute (GSI) of Japan has used a network of 200 continuously operating GPS receivers to monitor displacements associated with the earthquake zones that surround Japan.

On 28 December 1994, a magnitude 7.6 earthquake west of northern Honshu ruptured the fault that accommodates motion

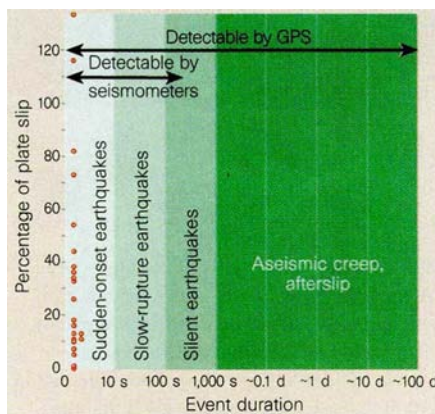


Figure 1 Processes that accommodate long-term motion of tectonic plates. Event duration is the time required for the majority of slip to occur; circles show slip released by sudden-onset earthquakes along different plate boundary segments as a fraction of the long-term plate slip (ref. 1). Along most plate boundaries, sudden-onset earthquakes account for less than half of the total slip since 1900. Measurements from the Global Positioning System (GPS) should in the future define the amount of slip accommodated over intervals longer than a few seconds.

between the subducting Pacific plate and overlying continental margin¹¹. The GPS array precisely recorded crustal displacements during the earthquake, and afterwards Heki and colleagues tracked the motions of 16 of the sites closest to the rupture zone. During the ensuing year, the post-rupture displacements of all but one of these sites exceeded their coseismic displacements, thereby indicating that significant earthquake afterslip had occurred.

No large earthquakes occurred along the fault plane during this period, implying that the afterslip was accommodated by aseismic creep. Modelling of the postseismic displacement vectors indicates that the energy released by the aseismic afterslip was comparable to that released during the earthquake, but was more evenly spread across the original rupture zone than was the coseismic motion.

These results have important implications for our understanding of the seismic cycle and of how faults accommodate plate motions. For example, large earthquakes along this part of the Japan Trench have peri-

odically relieved the strain that accumulates as the Pacific plate subducts beneath Japan. Historically, the energy released by these earthquakes has equalled only 20 per cent of the convergence rate (about 80 mm yr⁻¹), thereby raising the question of how the remaining 80 per cent occurs. Heki and colleagues demonstrate that significant aseismic slip can occur during the interval after an earthquake ruptures the asperities that are locking the subduction fault and before new asperities limit further motion. Other processes such as slow-rupture earthquakes accommodate additional slip^{4,5}; however, how much slip is unknown because slow-rupture earthquakes are largely undetectable by conventional seismometers. Continuous GPS observations should help to answer this question in the coming decades.

Heki and colleagues' results also add to the body of evidence that suggests that the concept of a 'stick-slip' seismic cycle is overly simplistic. Fault slip instead appears to vary in both space and time as a function of numerous variables including the geometric configuration of fault asperities, the frictional properties of materials bounding a fault, and changes in crustal stresses due to slip along nearby faults. Knowledge of fault slip during one seismic cycle may not necessarily imply predictability of fault slip in the future — which bodes ill for those who have hopes of earthquake prediction. At the least, many more observations are needed.

Over the past two-to-three years, continuously operating GPS receivers have been installed in actively deforming zones around the globe although none matches the impressive Japanese network employed by Heki and colleagues. Measurements from these receivers should provide many additional observations of fault slip before, during and after earthquakes. These will help to define whether significant aseismic afterslip is characteristic of interplate earthquakes or is instead peculiar to particular earthquakes and fault zones. In either case, our understanding of the manner in which the crust accommodates the unceasing motion of the tectonic plates will be greatly improved. □

Charles DeMets is in the Department of Geology and Geophysics, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.

1. Pacheco, J. E., Sykes, L. R. & Scholz, C. H. *J. Geophys. Res.* **98**, 14133–14159 (1993).
2. Solomon, S. C., Huang, P. Y. & Meinke, L. *Nature* **334**, 58–60 (1988).
3. Heki, K., Miyazaki, S. & Tsuji, H. *Nature* **386**, 595–598 (1997).
4. Kawasaki, I. et al. *J. Phys. Earth* **43**, 105–116 (1995).
5. Linde, A. T., Gladwin, M. T., Johnston, M. J. S., Gwyther, R. L. & Bilham, R. G. *Nature* **383**, 65–68 (1996).
6. Kanamori, H. & Kikuchi, M. *Nature* **361**, 714–716 (1993).
7. Beroza, G. C. & Jordan, T. H. *J. Geophys. Res.* **95**, 2485–2510 (1990).
8. Scholz, C. H., Wyss, M. & Smith, S. W. *J. Geophys. Res.* **74**, 2049–2069 (1969).
9. Bucknam, R. C., Plafker, G. & Sharp, R. V. *Geology* **6**, 170–173 (1978).
10. Marone, C. J., Scholz, C. H. & Bilham, R. *J. Geophys. Res.* **96**, 8441–8452 (1991).
11. Tanioka, Y., Ruff, L. & Satake, K. *Geophys. Res. Lett.* **23**, 1465–1468 (1996).

Where are the things we see?

Joel M. Miller and Christopher Bockisch

Our visual perception of the direction of objects seems to be based on several inaccurate visuomotor processes, which are the subject of two studies by Ross *et al.*¹ and Cai *et al.*² on pages 598 and 601 of this issue. Even a simple sideways glance to grasp a coffee mug involves visuomotor complexities that we have only begun to unravel. Suppose you flick your eyes to the mug: that is, you prepare (during a 200-millisecond latency period) and execute (another 50 milliseconds) a 'saccadic' eye movement. You note the location of the mug and the orientation of its handle, and look back, as you begin reaching for the mug — the whole process takes, perhaps, half a second. If you don't find yourself grabbing at air or mopping up spilt coffee then, supposing you moved your arm without continuous visual guidance, you have accurately assigned the information from your visual 'snapshot' of the mug to its proper relative location in space.

Locating visual targets would be simpler if our eyes did not move. An image falling on the specialized central fovea of the retina would then indicate an object lying straight ahead. An image falling to the left of fovea would correspond to an object to the right, and so on. But because only the fovea supports high-acuity trichromatic vision, foveate animals constantly scan visual scenes with saccades (brief, rapid movements), which quickly bring successive images to the fovea. However, a given retinal locus must then signify different spatial directions, depending on the position of the eye. The suddenness of saccades (which reach 400° per second in about 20 milliseconds) and their frequency (several each second) probably imposes an extreme computational burden, which can be used to study localization mechanisms.

We easily understand the spatial arrangements represented in cinematic images. Although we have no direct information about camera movement, patterns of visual flow can provide indirect information about continuous or overlapping camera movements, so that we can understand the visual worlds they represent³. Even 'jump cuts' between non-overlapping camera positions, which provide the viewer with no information about movement, do not seem unnatural.

It is possible that our visual world is assembled, at least in part, like a jigsaw puzzle, on the basis of visual cues within fragmentary images. But where visual snapshots do not overlap, there is no intrinsic basis for assembling them — that is, jigsaw-puzzle mechanisms do not work for large saccades (Fig. 1). Even for small saccades, visual-pro-

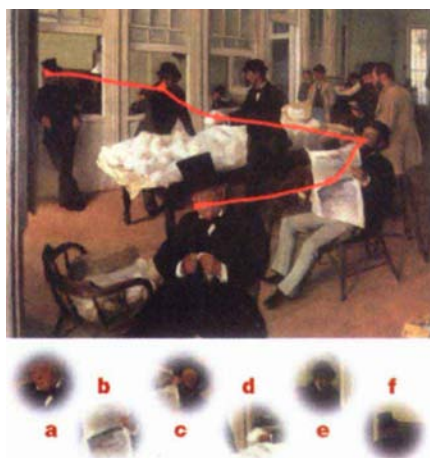


Figure 1 Detail of 'Portraits in an Office' (New Orleans) by Edgar Degas, with a hypothetical scanpath (red) showing six fixations joined by saccades. a–f indicate the predicted retinal content of these fixations. Based on information within the images, it seems possible that the visual system could properly relate the newspaper in b with its reader in c, and perhaps the cotton sample in d with the man to whom it is being presented in e. But information about eye position is probably required to ascertain that the man in f is watching the presentation, and that the man in a is self-absorbed.

cessing times imply that post-saccadic snapshots are not available for more than 50 milliseconds. In contrast, information about self-produced movement could be available for perception of direction even before a planned movement is completed. And it would be surprising if saccadic eye movements were not accompanied by information about 'camera' movement because, here, we are the cameramen.

Since at least the mid-nineteenth century, scientists have thought that the brain uses extraretinal (non-retinal) information about eye position (Fig. 2): knowing how the eyes have moved, the brain adds the angle of eye-rotation to the position on the retina of an object's image, to compute the spatial direction of the object^{4,5}. Several studies, placing retinal and extraretinal cues in conflict, have shown that rich visual information determines perceived changes in location, despite the contradictory extraretinal information⁶. But other experiments have shown the reverse⁷. Now, Cai *et al.*² describe a clear case in which extraretinal information is dominant, and Ross *et al.*¹ have mapped the time course of saccadic relocation, and show that there are combined effects of retinal and extraretinal information. In both studies, the measured saccadic relocation

was found to have systematic inaccuracies.

Cai *et al.*² exposed their subjects to a sparsely furnished visual world: there was nothing visible within 200 milliseconds of the subjects' saccades to a target, except for a vernier probe composed of a vertical pair of dots (which were visible from the start of the trial), and a briefly flashed dot (which was vertically centred between the flanking dots). The flashed dot preceded the movement of the eye to the target, at various horizontal locations in the neighbourhood of the saccade target, and subjects judged whether the flashed dot was collinear with the flanking dots. An accurate collinearity judgement might have been made from retinal-image information alone, during the brief period when flashed and flanking dots were simultaneously shown. Strikingly, in the 100-millisecond period before a saccade, flashes at all horizontal locations were mislocalized in the direction of the saccade, by about one-quarter the magnitude of the saccade. So visual direction is influenced by an extraretinal signal, even when more accurate, purely retinal, information is available. But there is a hint in the study of Cai *et al.* that the retinal information is also involved, because other studies⁸ involving only extraretinal information find even larger mislocalizations.

Ross *et al.*¹ asked subjects to make saccades from a fixation point to a target spot in a dimly lit environment, and to judge the horizontal position of a single vernier-probe flash relative to a ruler that was shown to them afterwards. In a second experiment, one component of a vernier probe was flashed early in the saccadic latency period (so as to be unaffected by relocation processes), and the other was flashed just before the saccade (when large localization errors are typically found). The authors found that visual objects presented in the 100-millisecond period around the beginning of a saccade are mislocalized by as much as half the magnitude of the saccade.

Ross *et al.*¹ and Cai *et al.*² have consolidated previous findings that probes around the fixation spot, and short of the saccade target, are mislocalized in the direction of the saccade. Ross *et al.* also show that probes which are presented beyond the target can be mislocalized in the opposite direction to the saccade. So, around the time of a saccade, the retinal direction-map not only translates — as it must to compensate for the movement of the eye — but also compresses in the vicinity of visual targets. That is, the retinal mechanisms, which are usually thought to be highly accurate, introduce a distortion here. In a third experiment, Ross *et al.* confirmed the compression effect by presenting horizontally spaced compound probes around the saccade target, and asking the subjects to judge them for separateness.

Although Cai *et al.* describe clear evidence for extraretinal influence on perceived direc-

tion (mislocalization of probes flashed beyond the saccade target, in the direction of the saccades), this is the opposite of the direction found by Ross *et al.* The studies straddle an interesting territory in which the visual system responds to both extraretinal and purely visual determinants of direction, and subtleties in the task determine the weight that is given to each. So, the study by Cai *et al.* shows (and that of Ross *et al.* is consistent with) imperfect compensation for eye movement, based on extraretinal information. The results from Ross *et al.* further indicate that visual objects, or the attention that is directed to them, can exert an attractive bias on the perceived locations of other objects. In both cases, the perceived directions of objects depend on complex interplay between many sources of flawed information.

Because vision is so important for effective behaviour, we would expect visuomotor mechanisms to exploit all possible sources of information about distal objects in a lighted environment. Distortion of biological signals, and the computational limitations of any single mechanism, may be overcome by using several mechanisms, to improve overall performance in a range of situations. The worst failings of such a 'patchwork' system can be minimized by still other mechanisms. For example, the visual instability that would be expected to result from transient mislocalizations around saccades is simply sup-

pressed in normal vision^{9,10}. Future studies in specially contrived visual situations could determine how several information sources interact in normal vision. This should allow improved design of, and performance in, artificial, high-information-flow environments, such as aircraft, in which, for example, suppressive mechanisms are undesirable. □

Joel M. Miller and Christopher Bockisch are at the Smith-Kettlewell Eye Research Institute, 2232 Webster Street, San Francisco, California 94115, USA.

Metrology

Stars and ships and superfluids

C. W. F. Everitt

In two elegant experiments, Schwab, Bruckner and Packard (on page 585 of this issue¹) and Avenel and Varoquaux², have measured the rotation of the Earth to 0.5% and 2% respectively by exploiting the macroscopic quantum properties of superfluid helium. Given the many current applications of gyroscopes, we may ponder the technical future of such exotic rotation sensors — and their relevance to ancient, still unsettling questions about absolute rotation.

Macroscopic quantization is a peculiar phenomenon in superconductors and superfluids, related to Bose-Einstein condensation but best understood by comparing the superconducting case with the de Broglie picture of the hydrogen atom. In hydrogen, stable electron orbits occur when the wavefunction of the lone electron closes on itself around the orbit. Fritz London conjectured³ in 1950 that an analogous condition holds in a superconductor, not for single electrons but for the collective behaviour of the roughly 10^{16} electrons (strictly, electron pairs) that constitute a circulating supercurrent.

This phenomenon has several intriguing consequences. One is that the magnetic flux through a superconductor is quantized in units of $hc/2e$ (h being Planck's constant, c the velocity of light and e the charge on the electron). Another, recognized in 1962 by Brian Josephson⁴, is that if one introduces one or more thin insulating layers (or other 'weak links') into a superconducting loop, there occur interference effects, resembling optical interference, which can be exploited to measure very accurately the magnetic field through the loop.

In superfluid helium it is vorticity rather than magnetic flux that is quantized, the unit of quantization being not $hc/2e$ but h/m , where m is the mass of the helium atom. An old experiment by G. B. Hess and W. M. Fairbank^{5,6} illustrates it. They cooled a tiny rotating bucket of liquid helium through its superfluid transition (2.16 K). Above 2.16 K,

1. Ross, J., Morrone, M. C. & Burr, D. C. *Nature* **386**, 598–601 (1997).
2. Cai, R. H., Pouget, A., Schlag-Rey, M. & Schlag, J. *Nature* **386**, 601–604 (1997).
3. Gibson, J. J. *The Ecological Approach to Visual Perception* 1–332 (Houghton Mifflin, Boston, 1979).
4. von Holst, E. *Br. J. Anim. Behav.* **2**, 89–94 (1954).
5. von Helmholtz, H. *Treatise on Physiological Optics* (Dover, 1925).
6. Matin, L., Picoult, E., Stevens, J., Edwards, M. & MacArthur, R. *Science* **216**, 198–201 (1982).
7. Rine, R. M. & Skavenski, A. A. *Vision Res.* **37**, 775–787 (1997).
8. Honda, H. *Vision Res.* **33**, 709–716 (1993).
9. Bridgeman, B., Hendry, D. & Stark, L. *Vision Res.* **15**, 719–722 (1975).
10. MacKay, D. M. in *Handbook of Sensory Physiology* 7, Part 3 (ed. Jung, R.) 307–331 (Springer, Berlin, 1973).

the helium behaves normally and is dragged viscously like water. For the superfluid, at slow rotations — where the total initial angular momentum is less than half h/m — the liquid stops rotating altogether, and the bucket correspondingly speeds up. Superfluid circulating through a thin hollow ring behaves similarly, but one can place a diaphragm with a minute hole in it (in practice, of diameter $\sim 1 \mu\text{m}$) across the section of the ring to form a 'weak link' in the superflow, producing a hydrodynamical Josephson effect. That is the key to the new rotation sensors. By modifying the topology of the ring to make the orifice part of a resonant chamber driven by a vibrating membrane, one can, as Schwab, Bruckner and Packard remark¹, determine "the state of absolute rotation of the containment vessel".

But what does that innocent phrase *absolute rotation* mean? We are drawn back to Newton's famous rotating bucket experiment. Centrifugal force makes a rotating liquid surface parabolic; so what is that rotation relative to? Not the bucket, for the surface can clearly be flat or parabolic with identical relative rotations. Newton concluded that it must be rotation with respect to 'absolute space'. Enter two centuries later Ernst Mach. He argued that inertia originates not from space but from the matter in space. According to Mach's principle, if we abolish from the Universe all matter except the bucket and water, the water will stay flat — centrifugal forces vanish.

Einstein made Mach's principle a founding assumption of General Relativity. He was, therefore, vastly annoyed when Kurt Gödel, having recovered from destroying mathematics, derived from General Relativity a most un-Machian model universe⁷. Gödel's universe is not ours, but that misses the point, which is that without special extra assumptions General Relativity does not explain inertia. One escape is *frame-dragging*. According to General Relativity, rotating massive bodies like the Earth partially

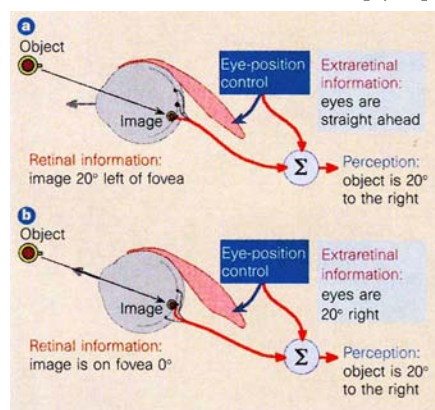


Figure 2 Simple scheme for combining retinal and extraretinal information on direction. The image of an object located 20° to the right is projected onto the retina, which signals the location of the image to a summing junction (retinal information). A copy of the command that determines eye position provides extraretinal information to the summing junction. The algebraic sum (Σ) appears at the output, and this is the perceived direction of the object. If retinal information accurately reflects the position of the image, and extraretinal information accurately reflects the position of the eye, then perceived direction is accurate, whether the eye is looking at the object (b) or elsewhere (a). But, as Ross *et al.*¹ and Cai *et al.*² now show, both sources of information are inaccurate around the time and destination of saccades.

drag space-time with them — an effect to be tested in NASA's orbiting gyroscope experiment, Gravity Probe B. Consider a gyroscope inside a massive rotating shell. It too should be dragged with the rotation, and if the whole Universe, assumed closed, is the shell, one can argue that the gyroscope will stay locked to the stars in Machian fashion. How difficult and shaky this argument is appears from the elaborate reasoning used to defend it in one well-known treatise⁸.

An exasperating problem is what the real issues are in Mach's principle. Take quantized rotation. The Hess–Fairbank experiment is a quantized Newton's bucket. When rotation stops, the controlling quantity is not mass but a quantum-mechanical wavefunction. Is this essentially connected with inertia? The condition for the appearance of vortices depends on h/m , involving mass, but the corresponding quantity for superconductors, $hc/2e$, does not even do that. Most of us would guess that frame-dragging affects all rotation sensors similarly, but some good theorists speculate that fundamental particle spins respond differently. We are a long way from full understanding.

It is time to descend from high theory to application. The word *gyroscope* (meaning 'to see circulation') was coined in 1851 by the French physicist J. B. L. Foucault, who in parallel with his famous pendulum also developed this device (a gimballed balanced fly-wheel) to demonstrate the Earth's rotation. These early instruments should not be underrated. In 1857 the mean error in determining the rate of rotation of the Earth from pendulum observations⁹ was less than 0.17%, and gyroscope performance was comparable.

With such brilliant results one might have expected a flood of applications, especially since gyroscopes drew the attention of such minds as Kelvin, Helmholtz and Maxwell. In fact, with one exception long antedating Foucault — the use of a spinning mirror as an artificial horizon for a sextant — little was done until 1898, when an Austrian engineer, Ludwig Obry, applied a quite simple gyroscope to perfect that beautiful, sinister invention, the Whitehead torpedo.

By far the most intellectually stimulating applications of gyroscopes hitherto have been in killing people. Within that melancholy truth the issues often differ from anything a physicist might imagine. For the torpedo, the problem was to increase the useful range of an object moving at 30 knots from a few hundred yards to a few miles. That only demanded a gyroscope capable of staying aligned to 0.5° for five minutes, far less than Foucault had achieved; the real difficulty lay in the control system, translating the gyroscope's directional information into steering. Even in the Second World War, the German V1 bombs, designed to hit central London, only needed gyroscopes accurate to 0.2°

per hour for 1.5 hours.

It is this interweaving of instrumental and operational issues that makes prediction in this field hard. Shortly after the Second World War, P. M. S. Blackett, a great physicist who had also been a naval officer and was one of the main inventors of operations research, wrote a masterfully wrong book. His *Fear, War and the Bomb* argued with brilliant cogency that the technical danger of atomic warfare was much exaggerated. No reasonable extrapolation of gyroscope performances over the next two decades would make accurate long-range ICBMs possible. How mistaken he was. The need created the advance.

Although performances of military gyroscopes remain classified, one may plausibly conjecture that 'compensated' drift-rates of the best case-rotated electrically suspended gyroscopes surpass the goal of a millionth of Earth's rotation which Schwab, Bruckner and Packard¹ have set for the next phase of their programme. One may also conjecture that drift performance is unlikely to be the determining factor in future military development.

Perhaps, with these and other new rotation sensors such as atomic interferometers¹⁰, we are returning to a situation where science rather than war drives invention. But physicists who enter this field would be wise to recognize and learn from the immense ingenuity that has been applied to gyroscope technology since 1945. □

C. W. F. Everitt is in the W. W. Hansen Experimental Physics Laboratory, Stanford University, Stanford, California 94305-4085, USA.

1. Schwab, K., Bruckner, N. & Packard, R. E. *Nature* **386**, 585–587 (1997).
2. Avenel, O. & Varoquaux, E. *Czech J. Phys.* **48**, Suppl. S6, 3319–3320 (1996).
3. London, F. *Macroscopic Theory of Superconductivity* 152 (Wiley, New York, 1950).
4. Josephson, B. D. *Phys. Rev. Lett.* **1**, 251–254 (1962).
5. Hess, G. B. & Fairbank, W. M. *Phys. Rev. Lett.* **19**, 216–219 (1967).
6. Hess, G. B. in *Near Zero: New Frontiers of Physics* (eds Fairbank, J. D., Deaver, B. S. Jr, Everitt, C. W. F. & Michelson, P. F.) 211–228 (Freeman, New York, 1988).
7. Gödel, K. *Rev. Mod. Phys.* **21**, 447–450 (1949).
8. Misner, C., Thorne, K. S. & Wheeler, J. A. *Gravitation* 546–547 (Freeman, New York, 1973).
9. Galbraith, J. A. & Haughton, S. W. *Manual of Astronomy* 2nd edn (Longman, London, 1857).
10. Gustavson, T. L., Bouyer, P. & Kasevich, M. A. *Phys. Rev. Lett.* **78**, 2046–2049 (1997).

Human artificial chromosomes

Providing a little stability

Peter E. Warburton and David Kipling

Molecular genetics and the explosion of information from the Human Genome Project will increasingly affect healthcare, as is already apparent in improved diagnostics and the understanding of many diseases. One exciting prospect is somatic gene therapy — the delivery (into cells) of DNA with therapeutic benefit; for example, the treatment of cancer or genetic disorders by replacing or correcting defective genes. But the functional elements of the main vehicle that is responsible for the organization, maintenance and inheritance of our genes — the chromosome — are at best poorly understood. This situation may now be on the mend, with a report from Hunt Willard's laboratory¹ in this month's issue of *Nature Genetics*. The authors describe the construction, and functional reintroduction into human cells, of the first mammalian artificial chromosomes (MACs). This success has implications not only for our understanding of chromosome function, but also in the development of MAC-based vectors for human gene therapy.

Current gene-therapy approaches are based on the delivery of therapeutic DNA, by packaging it in replication-defective recombinant viruses derived from, for example, retroviruses or adenovirus, or by the use of non-viral techniques, such as liposomes². All of these systems face, to varying degrees, problems of inefficient delivery to the target cell type, inappropriate patterns of gene

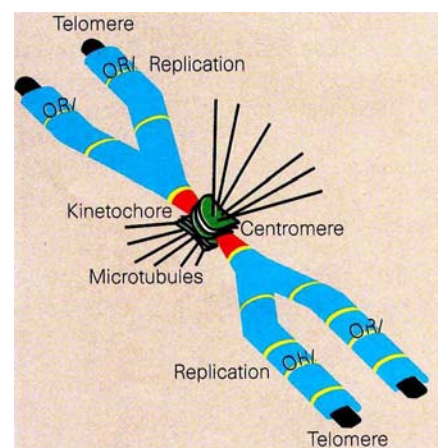


Figure 1 A linear chromosome, whether natural or a mammalian artificial chromosome (MAC), must provide for the replication of its DNA, telomeres to protect and replicate the ends of the chromosome, and a centromere to allow for segregation during cell division.

expression, and various toxic effects on cells². Moreover, immune reactions are possible, not only against viral proteins but also against the product of the therapeutic gene itself — because the gene is missing or mutated in the patient, the gene product may appear foreign to the immune system.

Most of these systems do not provide for long-term maintenance of the therapeutic DNA in the cell. In dividing cells, such as stem cells, long-term therapy requires repli-

cation of the introduced DNA and its segregation into daughter cells. At least two approaches are being explored towards this goal. The first uses plasmid vectors carrying parts of the Epstein–Barr-virus genome that allow the retention and replication of DNA under the control of the host cell-cycle³. Another is to re-create the natural cellular vehicle for DNA replication and segregation — the human chromosome.

The minimal requirements for an artificial chromosome include sequences to allow for the maintenance of its ends (telomeres), replication of its DNA, and mitotic segregation to daughter cells (centromeres) (Fig. 1). Yeast artificial chromosomes (YACs) have

been available for many years, and they function when introduced into cells as naked DNA. Mammalian geneticists, on the other hand, suffer from species envy, largely having had to be content with functional reintroduction of telomeres⁴. Moreover, whereas origins of replication are defined sequence elements in yeast, the existence of analogous mammalian origins remains controversial. One suggestion is that potential sequence-specific origins are frequent in the genome. Another is that replication simply requires a threshold amount of mammalian DNA, with the specific patterns of replication initiation that are observed being created by overall chromosome

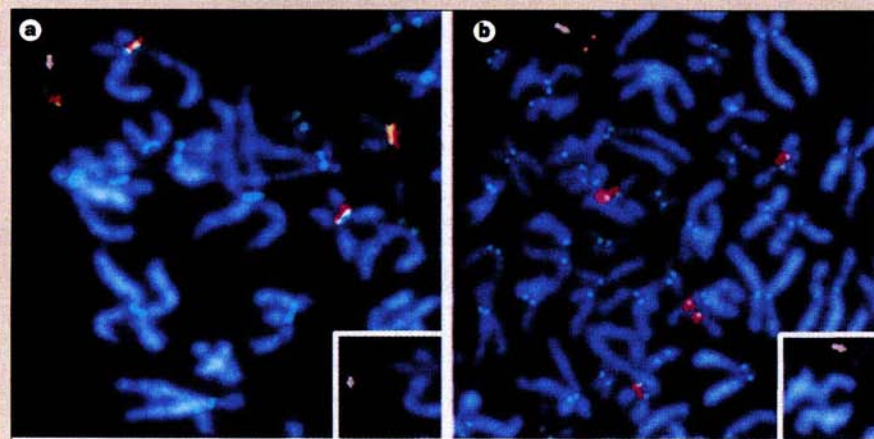
structure⁵. In either case, the replication of MACs may not require careful provision for origins.

The main stumbling block for the construction of MACs, which Harrington *et al.*¹ have now overcome, has been the provision of a centromere and the kinetochore that assembles there. The centromere is not simply a passive point at which the chromosome attaches to the microtubules of the mitotic spindle — it actively participates in the segregation of chromosomes to daughter cells. The centromeric region contains at least one motor, which is involved in chromosome movement during cell division. The centromere also mediates a critical checkpoint that prevents completion of cell division until all chromosomes are properly attached to the mitotic spindle⁶. The centromere (*CEN*) in *Saccharomyces cerevisiae* is a 125-base-pair DNA sequence; human centromeres, on the other hand, contain up to several million base-pairs of alpha satellite DNA, a large, tandemly repeating 171-bp DNA family⁷. Previous reintroduction studies had indicated that alpha satellite DNA could provide some aspects of centromere function^{8,9}, but the microchromosomes constructed by Harrington *et al.* (see panel) are the first to have all the features necessary to designate them MACs.

The successful creation of microchromosomes with centromere function from introduced naked DNA is a considerable step towards a MAC-based vector system for gene therapy. Some problems remain, such as the incompletely defined DNA that is used, and the variable and unpredictable structure, and relatively low yield, of the microchromosomes. This is further compounded by undesirable genomic rearrangements, which are caused by the introduced telomeric sequences. Ultimately, a practical MAC system for gene therapy must use defined DNA and yield microchromosomes in a predictable and efficient fashion. Even then, it still faces the problems of immune reactions against the therapeutic gene product, cell-specific targeting and so forth. Manipulation and delivery of such large DNA molecules is a technological challenge — these MACs are the whoppers of gene-therapy vectors — and it will be interesting to determine just how small a centromere-containing human chromosome can be. MACs do not have the stringent size limits of many viral-delivery systems, so they may allow for natural patterns of gene expression by introducing large genomic fragments carrying the therapeutic gene in its natural genomic environment.

Microchromosomes should also aid our understanding of chromosome maintenance and transmission in both mitosis and meiosis. For example, by custom-building arrays of mutated centromeric sequences from human and other species, the sequence

Ordering up a MAC



To make an artificial chromosome, Harrington *et al.*¹ obtained megabase-sized alpha satellite arrays, approaching the size of those at natural centromeres. Starting with a single characterized alpha satellite repeat unit, they used a directional-cloning method in a low-copy bacterial artificial chromosome vector, followed by *in vitro* ligation, to make tandem arrays up to 1 Mbp in size. This DNA was then mixed with human (TTAGGG)_n telomeric sequences and high molecular-weight genomic DNA, and introduced into cultured human cells using cationic liposomes. Although most clones represent integration or chromosome-fragmentation events, at

least two cases of *de novo* centromere formation seemed to occur on the introduced alpha satellite DNA. One such microchromosome, created from transfected DNA including alpha satellite from human chromosome 17, is shown. In most cells only a single microchromosome is seen, indicating faithful maintenance of copy-number control. These microchromosomes are stably inherited for many generations in the absence of selection, with a loss rate of less than 0.5 per cent per cell division. The formation of centromeres was further shown by the ability of the microchromosomes to bind two known centromere-associated proteins, CENP-C and CENP-E. **a**, The

microchromosome (arrow) was detected using antibodies against CENP-C (green), and the chromosome-17 alpha satellite was detected by *in situ* hybridization (red). **b**, A similar experiment using antibodies against CENP-E (green). Co-localization of green and red signals leads to a yellow signal. In both cases, the chromosomes were stained with the DNA dye DAPI (blue). These microchromosomes seem to be linear and they are 6–10 Mbp in size, but they are composed of a complex arrangement of sequences, characteristic of non-homologous recombination of the introduced DNA cocktail. (Figure reproduced courtesy of Harrington *et al.*¹). **P.E.W. & D.K.**

requirements for centromere function can be dissected. A likely place to start is the only conserved sequence between otherwise highly diverged mammalian centromeres — a 17-bp motif containing the binding site for the conserved centromere protein CENP-B (ref. 10). Microchromosomes could also allow other aspects of chromosome function to be explored, for example, the structure of, and regulation of the cell cycle by, the kinetochore. The development of mammalian artificial chromosomes as experimental and therapeutic reagents means that mammalian geneticists may soon be in for a big MAC attack. □

Peter E. Warburton is in the Institute of Cell and Molecular Biology, Michael Swann Building, King's Buildings, University of Edinburgh, Mayfield Road,

Edinburgh EH9 3JR, UK. David Kipling is in the Department of Pathology, University of Wales College of Medicine, Heath Park, Cardiff CF4 4XN, UK.

1. Harrington, J. J., Bokkelen, G. V., Mays, R. W., Gustashaw, K. & Willard, H. F. *Nature Genet.* 15, 345–355 (1997).
2. Crystal, R. G. *Science* 270, 404–410 (1995).
3. Calos, M. P. *Trends Genet.* 12, 463–466 (1996).
4. Kipling, D. *The Telomere* (Oxford Univ. Press, 1995).
5. Coverley, D. & Laskey, R. A. *Annu. Rev. Biochem.* 63, 745–776 (1994).
6. Pluta, A. F., Mackay, A. M., Ainsztein, A. M., Goldberg, I. G. & Earnshaw, W. C. *Science* 270, 1591–1594 (1995).
7. Tyler-Smith, C. & Willard, H. F. *Curr. Opin. Genet. Devel.* 3, 390–397 (1993).
8. Haaf, T., Warburton, P. E. & Willard, H. F. *Cell* 70, 681–696 (1992).
9. Larin, Z., Fricker, M. D. & Tyler-Smith, C. *Hum. Mol. Genet.* 3, 689–695 (1994).
10. Kipling, D. & Warburton, P. E. *Trends Genet.* 13, 141–145 (1997).

Animal behaviour

The monkeys' defence alliance

Robin Dunbar

Social life remains one of the most obvious characteristics of the living world. Yet it has not always been clear why animals should live in groups. The costs of sociality are often considerable: they can involve both direct effects through competition over ecological resources¹ and indirect effects such as the increase in infertility caused by increased stress². The gradient created by these costs demands significant advantages if natural selection is to overcome these disadvantages in order to make group-living worth any animal's while.

Conventional wisdom identifies a number of possible selective advantages to group living, including cooperation in rearing (as in termites and many social insects, many birds and at least some members of the dog family) and hunting (lions, hunting dogs and hyaenas, for example). These advantages do not, in general, apply to large mammals such as primates which neither hunt nor, with very limited exceptions, rear offspring

communally. Instead, primates are thought to be social either to defend themselves against predators or to defend access to food resources¹. Although several attempts have been made to test these competing hypotheses^{3–5}, the issue has remained obdurately unresolved and much disputed^{6,7}. In a paper in *Proceedings of the Royal Society*⁸, however, Ronald Noë and Redouan Bshary now provide clear-cut support in favour of one of these hypotheses.

The resource-defence hypothesis owes its origin to the observation that reports of predation in primates are in fact rare, whereas disputes over territorial boundaries (and sometimes direct access to food sources) are common. Wrangham⁹ argued persuasively that food-resource defence was thus more likely to be the prime mover in promoting sociality among primates. The debate has swung back and forth for more than a decade, in part because most attempts to test between these views have tended to focus on one with-

out taking account of the other. Too often, the competing hypotheses were confounded; as a result, many studies came to opposite conclusions from similar starting points.

It has also become apparent that there is a deeper underlying problem with the logic of the earlier arguments. We cannot assume that an absence of mortality due to predation means that predation pressure is not a significant factor in the lives of animals. We should not expect species as behaviourally flexible as primates, for example, to continue to pursue some species-typical pattern of behaviour while accepting high levels of mortality from predators as some kind of inevitability. If group-living is an adaptive response to predation pressure, then we ought to expect primates (and possibly many other species) to adjust their group sizes so as to minimize this pressure¹⁰. In other words, what we see in terms of actual predation rate is not the true predation risk experienced by the animal, but what is left over as uncontrollable risk once the animals have done what they can to minimize actual mortality.

Such a view encourages a number of alternative approaches. One is to study the animals' responses to predation pressure; another is to stimulate these responses experimentally. If predation risk is indeed a significant factor, then we should see an immediate response by the animals to the challenge offered by predators. In contrast, if primates form groups for reasons related to resource defence, then we should see no response to such challenges, and primates should ignore experimental evidence of the proximity of predators. It is these options that Noë and Bshary⁸ have exploited in an elegant series of field experiments.

During a large-scale study of monkey ecology in the Tai National Park in the Ivory Coast, West Africa, Noë and Bshary were able to capitalize on a tendency for many tropical-forest primate communities to form mixed-species foraging parties, a phenomenon that others have suggested might be a response to predation risk¹¹. In their study area, two species (red colobus and diana monkeys) showed a particularly strong affinity for each other. Both species (but particularly the red colobus) suffer from predation by chimpanzees in many parts of Africa^{12–14}. However, chimpanzee hunting is a strikingly seasonal activity in the Tai, being almost entirely confined to the early dry season each year, when the chimpanzees form large enough groups to facilitate cooperative hunts. At this time of year, the Tai colobus groups experience roughly 0.7 chimpanzee attacks per month, with a mean kill rate of one animal per two attacks.

Noë and Bshary took the opportunity provided by this natural experiment to examine the frequencies with which the two species formed mixed-species parties in different seasons in each of four years for two

RONALD NOË

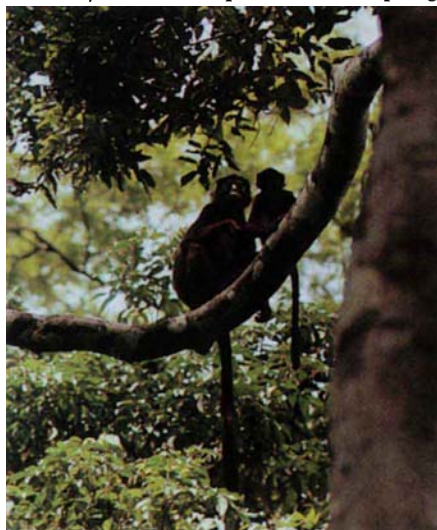


Figure 1 Alone but united — members of the groups studied by Noë and colleagues, red colobus (left) and diana monkeys.

different pairs of associated colobus–diana groups. Association rates were higher in the chimpanzee hunting season than during the other three seasons of the year in six out of eight cases (when the null hypothesis would predict only two would be higher). The more vulnerable red colobus were responsible for initiating mixed foraging parties significantly more often than the diana monkeys.

This effect was then confirmed by an experimental study in which the authors played tape recordings of the loud calls of chimpanzees from hidden speakers. They showed that, when the two species had spent the night together, a short bout of chimpanzee calls broadcast in the early morning was significantly more likely to prolong the association between the two species during the day than a bout of leopard calls (a less serious predator for the monkeys), empty tape or the noise of a generator.

They then repeated the experimental design in reverse, and showed that chimpanzee calls were significantly more likely than any of the other three auditory stimuli to result in the colobus and diana groups converging to form a mixed-species party when they had been separated by more than 100 metres.

This, then, seems to clinch the matter. These primates appear to use grouping as a deliberate strategy for minimizing their risk of attack by hunting chimpanzees. Mixed-species parties have an ecological advantage

over single-species parties of comparable size because the different foraging niches of the two species reduce the level of ecological competition experienced by the members of supergroup¹⁵. One final point remains to be confirmed, however: are chimp predation rates actually lowered by the formation of mixed-species parties? □

Robin Dunbar is in the ESRC Research Centre in Economic Learning and Social Evolution, Department of Psychology, University of Liverpool, Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK.

1. Janson, C. I. in *Evolutionary Ecology and Human Behaviour* (eds Smith, E. A. & Winterhalder, B.) 95–130 (Aldine de Gruyter, New York, 1992).
2. Bowman, L. A., Dille, S. R. & Keverne, E. B. *Nature* **275**, 56–58 (1978).
3. van Schaik, C. P. *Behaviour* **87**, 120–144 (1983).
4. Cheney, D. & Seyfarth, R. M. *Behav. Ecol. Sociobiol.* **21**, 375–386 (1987).
5. Barton, R. B., Byrne, R. B. & Whiten, A. *Behav. Ecol. Sociobiol.* **38**, 321–329 (1996).
6. Isbell, L. A. *Evol. Anthropol.* **3**, 61–71 (1994).
7. Boinski, S. & Chapman, C. A. *Evol. Anthropol.* **4**, 1–3 (1995).
8. Noe, R. & Bshary, R. *Proc. R. Soc. Lond. B* **264**, 253–259 (1997).
9. Wrangham, R. W. in *Primate Societies* (eds Smuts, B. B., Cheney, D. L., Seyfarth, R. M., Wrangham, R. W. & Struchiner, T. J.) 282–296 (Univ. Chicago Press, 1987).
10. Hill, R. A. & Dunbar, R. I. M. *Behaviour* (in the press).
11. Gautier-Hion, A., Quir, R. & Gautier, J.-P. *Behav. Ecol. Sociobiol.* **12**, 325–335 (1983).
12. Boesch, C. & Boesch, H. *Am. J. Phys. Anthropol.* **78**, 547–573 (1989).
13. Stanford, C. *Anim. Behav.* **49**, 577–587 (1995).
14. Basabose, K. & Yamagiwa, I. *Primates* **38**, 45–55 (1997).
15. Wachter, B., Schabel, M. & Noe, R. *Ethology* (in the press).

Exobiology

A lively debate

Timothy J. McCoy

In August last year, D. S. McKay *et al.* (*Science* **273**, 924–930; 1996) described evidence from a meteorite that life might once have existed on Mars. So are we close to answering the old question of whether there is life outside the Earth? A conference last month* gave us a chance to debate the latest evidence.

Only in the past 15 years has the martian origin of a certain rare group of meteorites been accepted. The life-on-Mars debate centres on just one of them, the 4.5-billion-year-old orthopyroxenite Allan Hills (ALH) 84001, because it alone is old enough to have experienced the warm, wet, early martian climate that has been postulated. It was probably during this period that fluids permeated ALH84001, forming the roughly one per cent by volume of carbonates now found in the rock. McKay *et al.* argued that a variety of chemical, mineralogical and morphological features of these carbonate globules, taken together, indicated that they had been deposited by microbial life on Mars.

Terrestrial bacteria are not known to



Figure 1 Rock of uncertainty. A 212-gram fragment of Allan Hills 84001, the meteorite that holds controversial evidence of past life on Mars.

survive above about 120 °C, so mineral thermometry, which suggests that the meteorite's carbonates formed at temperatures of up to 700 °C, does not favour the presence of life (R. P. Harvey and H. Y. McSweeney Jr *Nature* **382**, 49–51; 1996). But the complex chemical zoning within the magnesium, iron and calcium carbonates may be inconsistent

with equilibrium formation, shedding doubt on such thermometry (G. A. McKay, NASA/Johnson Space Center; A. H. Treiman, Lunar and Planetary Inst.). Evidence instead for a low-temperature origin comes from variable natural remnant magnetic orientations, suggesting carbonate formation below the magnetic closure temperature of pyrrhotite (perhaps less than 100 °C) (J. L. Kirschvink, California Inst. Technol.).

The distribution of oxygen isotopes in the meteorite is also uneven. J. W. Valley (Univ. Wisconsin) argued that this isotopic disequilibrium, primarily by analogy with similar systems on Earth, favours carbonate formation below 100 °C. But the correlation between major-element and oxygen-isotope compositions implies that the carbonates formed either at high temperature with a limited supply of carbon-dioxide-rich fluid; or in a cooling, open system with abundant water-rich fluid (L. A. Leshin, Univ. California at Los Angeles; J. D. Gilmour, Univ. Manchester). Even that low-temperature model requires cooling over a range of about 200 °C and peak temperatures beyond those amenable to terrestrial bacteria.

Two speakers (G. A. McKay; E. R. D. Scott, Univ. Hawaii) presented striking evidence that several phases — perhaps even the carbonate globules — had been mobilized by shock after carbonate formation. The effects of shock and mixing on all these investigations are not fully appreciated, and they may explain some of the apparent contradictions between different studies.

Sulphur isotopes are of particular interest in the debate, because they are highly fractionated by most terrestrial bacteria. Sulphide-rich carbonate areas in ALH84001 have compositions that are indistinguishable from large magmatic sulphides elsewhere in ALH84001 and in other martian meteorites (J. P. Greenwood, Univ. Tennessee; C. K. Shearer, Univ. New Mexico), and although the analysis of nanometre-sized sulphide grains within a carbonate matrix is difficult, their unfractionated compositions cast considerable doubt on a biological origin.

Another strand of the original evidence was the appearance of 'nanofossils'. Acid etching of ALH84001 has now revealed structures that can be interpreted as bacterial secretions (D. S. McKay, NASA/Johnson Space Center). The mineralogy and composition of these materials remains uncertain, however, and so their origin is ambiguous. Atomic force microscopy provided direct examination of the nanofossils (A. Steele, Univ. Portsmouth): the elongated and ovoid features are found on fresh, untreated and uncoated samples, and so they are not artefacts of sample preparation (as had been claimed). But the occurrence of these biofilms and nanofossils, as well as organic molecules (G. Flynn, SUNY-Plattsburgh) in parts of the meteorite not associated with carbonates,

* Lunar and Planetary Science Conference, Houston, Texas, USA, 17–21 March 1997.

might imply an origin by weathering or terrestrial contamination.

Perhaps the most hotly debated topic of the conference was the nature of the magnetite crystals found within the carbonates, and the implications for the origin of the nanofossils. Magnetites have been found in the meteorite that are a similar shape to the McKay *et al.* nanofossils, hinting that the nanofossils may be non-biogenic magnetite (J. P. Bradley, MVA, Inc.). They contain screw dislocations, which are inconsistent with low-temperature formation and have not been observed in biogenic magnetites. Another study (K. Thomas-Keperta, Lockheed-Martin) failed to see such elongated forms, but did find teardrop-shaped magnetites only previously found in biological systems. The source of the discrepancy between these studies is unclear, but may lie in differences in sample preparation.

Few terrestrial rocks have been examined on the nanometre scale for evidence of microbial activity, although that is slowly changing. Columbia River basalts contain

bacteria with appendages that are similar in shape and size to those in ALH84001 (Thomas-Keperta). And both magnetite and carbonate can be made by bacteria in laboratory experiments (H. Vali, McGill Univ.).

This conference saw the first round of discussions about past biological activity on Mars. An emerging opinion was that the carbonates could have formed through low-temperature, non-biological processes, offering an alternative to the older polarized views of 'hot and lifeless' versus 'cold and alive'. Many participants, even those who disagree with the conclusions of McKay *et al.*, believe that life may have once existed on Mars and that the search for evidence of it should continue. This conference will clearly not be the last word — proposals for new allocations of ALH84001, and for the necessary funding, are already being considered. Soon, we may know whether we once shared our Solar System with other life forms. □

Timothy J. McCoy is in the Department of Mineral Sciences, MRC 119, Smithsonian Institution, Washington DC 20560, USA.

Seismology

Tracking slabs in the lower mantle

Guy Masters

Sometimes it seems as if geoscientists as a group are exceptionally bad at solving problems. For example, for about 30 years we have been debating whether or not there is a significant barrier to mantle flow at a level close to the depth of the 660-km seismic discontinuity within the Earth. Our whole understanding of the thermal and chemical evolution of the Earth hinges on this fundamental issue. Most geophysicists and geodynamists lean towards the view that there is significant mass transfer between the 'lower' and 'upper' mantles. Many geochemists firmly believe the opposite must be true. Perhaps the tomographic images presented in the paper by van der Hilst *et al.* (beginning on page 578 of this issue¹) will help the geochemists to change their minds.

Several factors make this new contribution notable. The first is that van der Hilst and colleagues' images clearly show narrow compressional velocity anomalies in the lower mantle that are geometrically related to the well-known seismogenic subduction zones in the upper mantle. Such images make it hard to avoid the conclusion that there is significant slab penetration into the lower mantle. The authors argue that it is the careful reprocessing of the basic travel-time data which makes these new images possible.

The second factor is the close correlation of the model of compressional velocity anomalies with the independently determined shear velocity model of Grand^{2,3} (compressional waves are the fastest of seis-

mic waves, and are known as P (for primary); shear (S for secondary) waves are slower). Grand's model is based not only on completely different data but also on different processing techniques, so the correspondence between the two models is convincing. Furthermore, the location of the fast anomalies correlates quite well with reconstructions of slab geometry inferred from the recent history of subduction. In fact, the authors suggest that some of the discrepancies between the seismic images and the slab models can be used to refine our knowledge of subduction history.

So how might sceptical geochemists wiggle out of this problem? First, they could appeal to the approximate nature of the algorithm used for converting the travel-time data into an image of velocity anomalies. This particular algorithm might lead to the elongation of structures or even the creation of artefacts. However, comparison of the new images with more conventional global tomographic models⁴ of structure at longer wavelengths shows that the same basic features are apparent. (This is particularly true of a recent compressional velocity model based on hand-picked, long-period travel times made from modern digital data⁵.)

Second, they could point out that many seismologists are themselves sceptical about the exact width of the features being imaged and, in particular, the amplitudes of the anomalies (which are generally underestimated in tomographic inversions of this

type). For example, there is good evidence that the relative perturbation in shear velocity is about 1.7 times the relative perturbation in compressional velocity in the mid lower-mantle (consistent with a thermal origin for the anomalies), whereas the images presented in this issue show the compressional and shear velocity anomalies being of similar size. Obviously, further refinement of the images is needed and the velocity anomalies in the lower mantle will change their amplitude and shape somewhat — but they are not going to disappear.

On the other hand, geochemists have quite compelling reasons for believing that a significant fraction of the mantle is in a chemically primitive state, indicating that the mantle is at least partly segregated. For example, the atmosphere contains roughly half of the mass of ⁴⁰Ar created by the decay of ⁴⁰K over Earth history and a reasonable place to put the missing argon is in an isolated lower mantle. Similarly, the flux of ⁴He observed at ocean ridges is only a small fraction of that expected from radiogenic heat production within the mantle — an isolated lower mantle is again a convenient place for the rest. Other lines of evidence point in this direction and Hoffman⁶, in a review published earlier this year, concludes that the geochemical evidence still favours some kind of mantle layering, but one in which the layers are not completely isolated.

Can a model which reconciles all the geophysical and geochemical observations be constructed? Many solutions have been suggested but, in the absence of direct observational constraints on the mass flux between the upper and lower mantles, we cannot get a quantitative answer. The images shown in the paper by van der Hilst *et al.* begin to provide these much-needed constraints.

Hoffman concludes his review of mantle geochemistry by saying: "The controversy over the issue of convective layering and the depth of plume sources is likely to be resolved when the so far rather fuzzy seismic imaging of the Earth's interior becomes comparable in resolution to that achieved by geochemical mapping". Even if we accept this rather provocative characterization of the relative merits of seismological and geochemical mapping, I think we can all agree that the images shown in Fig. 3 of the paper (page 581) are far from fuzzy. □

Guy Masters is in the Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California 92093-0225, USA.

1. van der Hilst, R. D., Widiyantoro, S. & Engdahl, E. R. *Nature* **386**, 578–584 (1997).
2. Grand, S. P. *J. Geophys. Res.* **99**, 11591–11621 (1994).
3. Grand, S. P., van der Hilst, R. D. & Widiyantoro, S. *Geol. Soc. Am. Today* (in the press).
4. Su, W.-J., Woodward, R. L. & Dziewonski, A. M. *J. Geophys. Res.* **99**, 6945–6980 (1994).
5. Bolton, H. thesis, Univ. California San Diego (1996).
6. Hoffman, A. *Nature* **385**, 219–229 (1997).

Space and time in the mental universe

Patricia Goldman-Rakic

The term 'working memory' refers to the ability to bring to mind and process limited amounts of information. The defining quality of working memory is its transient, 'on line', nature: for example, holding and processing pieces of information as we read and comprehend a line of text, recall a telephone number, plan a chess move or follow verbal instructions to locate an unfamiliar street address. Working memory provides a temporal bridge between events — both those that are internally generated and environmentally presented — thereby conferring a sense of unity and continuity to conscious experience. Advances in human brain imaging have opened a window on the anatomy of higher cognitive functions¹ such as working memory, and magnetic resonance imaging (MRI) techniques have now refined our ability to observe the timing and sequences of such functions. This is exemplified by Cohen *et al.*² and Courtney *et al.*³ (pages 604 and 608 of this issue), who have looked at the temporal parameters of working memory in the human brain.

Current understanding of working memory in the central nervous system is based, in part, on psychological studies of short-term memory in humans^{4,5}, and neurobiological

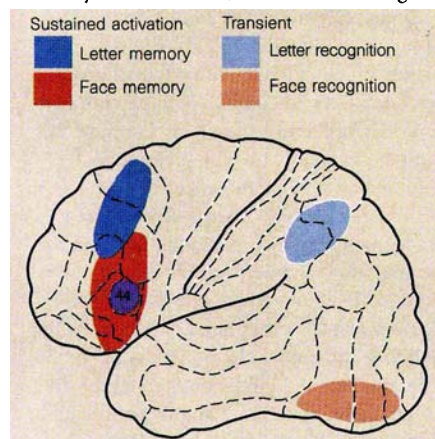


Figure 1 Sophisticated brain-imaging techniques have allowed the areas that are associated with human working memory to be identified. Cohen *et al.*² asked subjects to remember whether a given letter was the same as a letter that had been presented two or three trials back in a series. Courtney *et al.*³ set subjects a similar task that required them to recognize and remember particular faces. Specific regions of the prefrontal cortex, along with functionally related (and presumably anatomically connected) areas in the sensory-related regions of the posterior lobe, were identified as being active in each working-memory task.

analyses of the same in non-human primates^{6,7}. Single-neuron recordings in monkeys that are trained to perform working-memory tasks have identified components of a working-memory circuit in the prefrontal cortex. Here, the neuronal processes that are related to task performance can be dissociated, on the scale of milliseconds to seconds^{6,7}. So, during a working-memory task, as the stimulus is sequentially registered, stored over a period of seconds and then translated into a motor response, specific neuronal populations respond in characteristic ways. For example, one class of prefrontal neuron responds to a visual stimulus as long as the stimulus is in view. In contrast, other prefrontal neurons are activated at the offset of the stimulus, and they remain active (at heightened levels) all the time that the monkey has to remember the location^{6,7}, or features, of an object⁸. These neurons are thought to be the cellular correlate of the transient memory trace.

Cohen *et al.*² and Courtney *et al.*³ show that the fast scan-times of imaging methods such as echoplanar functional magnetic resonance imaging (fMRI) can accurately capture the time course of the analogous human working memory. They extend the knowledge that has been gained from animal models by single-cell recording and other techniques, to show surprisingly similar phenomena in normal humans.

Functional MRI scans detect changes in blood flow and volume, which reflect local increases in brain activity. Scientists have now begun to take advantage of the temporal resolution that is possible with MRI to discriminate between different stages in cognitive processing. When the spatial resolution of MRI technology is combined with the temporal aspect, specific functions and subfunctions can be related to distinct anatomical structures in the human brain, allowing a number of unresolved issues about the functional architecture of cognition to be addressed.

One such issue concerns whether the cognitive functions of the prefrontal cortex are organized hierarchically — that is, does information processing become progressively centralized through successive stages of information flow? For example, sensory systems are anatomically segregated at early stages of cortical processing into visual, auditory and somatosensory cortical areas, and their associated pathways. According to the accepted hierarchical model of cognition, the working-memory system has two components: a 'central executive', which is non-



100 YEARS AGO

The ostrich, *Struthio camelus*, has been observed with interest from very early times; it has frequently been the subject of remark by African travellers; and it has been domesticated and farmed in the Cape Colony for some thirty years. Yet it is remarkable how little is known about it in scientific circles, and how many misconceptions still prevail as to its nature and habits. This article is founded on personal observations made during nine years of uninterrupted ostrich-farming in the Karroo of the Cape Colony, and during travels about the country generally. In large ostrich camps, some of which are a couple of miles in diameter, numbers of birds of both sexes run in what is practically a wild state, seldom interfered with in any way, except when rounded up to be plucked or to be fed in a drought. The habits of birds thus farmed differ in no way from those of native wild birds, except perhaps that monogamy is more difficult. ... The ostrich hen lays every other day, and the egg weighs about three pounds; it is a tasty and nutritious food however prepared, very rich, and excellent for making pastry and cakes. It is generally computed to be equal to two dozen fowls' eggs; but this must be on account of its superior richness, for, from personal experiment, the empty shell of a fairly large one exactly held the contents of eighteen fowls' eggs. It takes about forty minutes to boil an ostrich egg hard. The period of incubation is about six weeks.

From *Nature* 8 April 1897.

50 YEARS AGO

An account has been received from Mr. S. Duvdevani, of the Palestine Meteorological Service, of an optical method of estimation of dew which he has developed at the Dew Research Station, Pardess Hanna Agricultural School, Karkur, Palestine. The need was felt for a much quicker and less costly method than that of direct weighing of the dew collected on a standard surface, so that a network of voluntary observing stations could be organised for a country-wide study of the amount of dew in the different seasons. In a dry climate such as that of Palestine, the success of certain crops is largely dependent upon the amount of dew...

From *Nature* 12 April 1947.

denominational, and two 'slave systems'. The slave systems store visual and spatial information ('visuo-spatial sketchpad') or auditory information ('phonological loop'), and they keep it activated ('on line')⁹ for subsequent processing by the central executive. The anatomy of the central executive and slave systems is not well understood but, consistent with their distinct cognitive functions, they are thought to be anatomically segregated, with the central executive in the prefrontal cortex, and the slave systems associated with posterior cortical regions (for example, the parietal and temporal cortices).

Now, Courtney *et al.*³ have found that at least some functions of the slave systems may be carried out in the human prefrontal cortex. They presented subjects with pictures of human faces, and asked them to recall whether the picture being shown was the same, or different, from one that had been presented eight seconds earlier. The sequence of events in the task, and the time series used, allowed the authors to work out whether the activation of neurons in different regions correlated with the sequence of visual stimulation *per se*; pictures of faces in particular; or the interpolated delay periods between the presentation of stimuli (when the sequentially presented stimuli were being held 'on line'). The authors found that activations in the prefrontal areas correlated most strongly with delay periods, compared with activations in the visual areas, which were more strongly correlated with sensory stimulation. Notably, the working-memory task that was used by Courtney *et al.* mainly activated different areas of the human prefrontal cortex from those that were activated in the study by Cohen *et al.*² (Fig. 1). The one major area of overlap seems to be area 44, which is part of the rostral language zone (Broca's area). As area 44 is common to both tasks, it could be argued that it represents the central executive in the frontal lobe. Alternatively, the overlap could indicate that the subjects used verbal mediation to bridge the temporal delays in the two different tasks.

Cohen *et al.* presented subjects with written consonants, one at a time every ten seconds, and asked them to judge whether each consonant was the same as a letter presented one, two or three trials back in the sequence. This is known as the 'n-back' task¹⁰, and it requires that subjects remember the order of consonants, as well as their identity. The farther back in the sequence the consonant to be recalled occurs, the greater the 'load' on working memory. Following the presentation of each stimulus, many scans were acquired to track the sequence of brain activations in different brain regions as the subjects processed the information in working memory — much as single-unit recording studies track neuronal activity in monkeys during the delay in a delayed-response trial.

Cohen *et al.* show that activations in the

prefrontal cortex are maintained throughout the ten-second interstimulus interval and, importantly, that the degree of prefrontal activation is higher for the conditions with the greatest memory load. By contrast, activations in the primary visual, somatosensory and motor cortices, as well as in several secondary sensory regions, are not sustained across the ten-second 'memory' interval, and they are not related to memory demand. These areas are probably responsive to the sensory or perceptual, but not the mnemonic (memory-aiding), aspects of the working-memory task. Because consonants presented two or three trials back in the sequence require a certain amount of processing, as well as maintenance of information (for example, updating the order and content of working memory), the new results², together with previous studies¹¹, indicate that the central-executive component may also be located in prefrontal areas.

The findings of Cohen *et al.*² and Courtney *et al.*³ are in close harmony with a broad basic science and clinical literature which segregate perceptual and mnemonic functions in anterior and posterior association cortices, respectively. The powerful methods used in these studies now provide a real-time analysis of the temporal dynamics of working-memory functions, in combination with brain imaging. Armed with this information about the locus and sequencing of working-memory processes which are thought to be the domain of slave systems, the next step will be to search for the elusive central executive. Given the increasingly powerful techniques available, including electroencephalography¹⁰ and magnetic source imaging¹² (which both have temporal resolutions in the millisecond range), we are ushering in a new era for understanding the cognitive processes that define human mental action. □

Patricia Goldman-Rakic is in the Section of Neurobiology, Yale University School of Medicine, 333 Cedar Street, New Haven, Connecticut 06510, USA.

1. Raichle, M. *Cereb. Cortex* **6**, iii–iv (1996).
2. Cohen, J. D. *et al. Nature* **386**, 604–608 (1997).
3. Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. *Nature* **386**, 608–611 (1997).
4. Baddeley, A. & Hitch, G. in *The Psychology of Learning and Motivation: Advances in Research and Theory* (ed. Bower, G. H.) 47–89 (Academic, New York, 1974).
5. Carpenter, P. A. & Just, M. A. in *Complex Information Processing: The Impact of Herbert A. Simon* (eds Klahr, D. & Kotovsky, K.) (Erlbaum, Hillsdale, NJ, 1989).
6. Goldman-Rakic, P. S. in *Handbook of Physiology, the Nervous System, Higher Functions of the Brain* (ed. Plum, F.) 373–417 (Am. Physiol. Soc., Bethesda, MD, 1987).
7. Fuster, J. M. *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate* (MIT Press, Cambridge, MA, 1995).
8. Wilson, F. A. W., O'Scalaidhe, S. P. & Goldman-Rakic, P. S. *Science* **260**, 1955–1958 (1993).
9. Baddeley, A. *Working Memory* (Oxford Univ. Press, London, 1986).
10. Gevins, A. S. *et al. Electroenceph. Clin. Neurophysiol.* **76**, 339–350 (1990).
11. D'Esposito, M. *et al. Nature* **378**, 279–281 (1995).
12. Usitilo, M. A., Williamson, S. J. & Seppa, M. T. *Neurosci. Lett.* **213**, 149–152 (1996).

Daedalus

Mutable mutations

Antibiotics, those former heroes of medicine, are fast losing their power. Diseases such as tuberculosis, pneumonia and gonorrhoea, once easily cured by them, are creeping back to haunt us. We all know why. Antibiotics have been dispensed liberally and carelessly; many are routinely spread around in animal feeds. As a result, many bacteria have encountered them at low levels over long periods of time. They have evolved mutations which resist them. The worst offenders are 'hypermutable' bacteria, whose mechanisms for avoiding mutation have themselves mutated.

So Daedalus wants pharmacologists to do some mutating of their own. He points out that antibiotics are fungal metabolites or derivatives of them. The fungi — typically moulds — evolved them to discourage competing bacteria, in a chemical contest which has been going on for millions of years. Each time the fungi improved their brand of penicillin or streptomycin, the bacteria slowly mutated to become resistant to it. The fungi then had to develop an even deadlier version. The time is ripe for them to take the next step.

DREADCO biochemists are now helping them. They are putting the most accomplished moulds and the most virulent, hypermutable bacteria together in the same culture. Its medium is dense with mutagens, and irradiated from a radioactive source. The conditions will be adjusted to put the fungi at a slight disadvantage. This fierce and continuous selection pressure will force hypermutation in both fungi and bacteria. A year of such ruthless competition will advance the antibiotic art by many centuries.

The DREADCO team will then face a tricky problem. The mutated fungi will be immensely valuable, a source of a whole new stable of powerful antibiotics. The bacteria, on the other hand, will be deadlier and more antibiotic-resistant than ever. The biochemists will have to perfect some sort of filtering or centrifuging process which will permit the fungi to be harvested while reliably destroying every single one of the bacteria. DREADCO will then have a whole new range of antibiotic fungal metabolites. The old medical magic will return. But will we have learned from past experience? Daedalus is confident that his new antibiotics will once more be sprayed freely around, carelessly over-prescribed, and incorporated into numerous bulk non-medical products. In due course they will lose their potency, and DREADCO will be able to work the same trick again.

David Jones

Ammonia excretion by hummingbirds

There are three main excretory end-products of nitrogen metabolism in vertebrates, and their relative proportions in the urine reflect both phylogeny and environment¹. Fishes, amphibians, and aquatic reptiles predominantly excrete ammonia (ammonotelic) or urea (ureotelic), whereas terrestrial reptiles have adopted uric acid as a more water-efficient means of excreting nitrogen (uricotelic). Uric acid is also thought to be the predominant end-product in birds, but nitrogen excretion has never been examined for birds with very high rates of water flux. Here we report the first demonstration of ammonotelic in a bird, the Anna's hummingbird (*Calypte anna*), and present evidence that this nectarivore is ammonotelic under some conditions and uricotelic under others.

Hummingbirds have the highest metabolic rates of any vertebrate², and their energy requirements are met largely by the consumption of floral nectar. At low ambient temperatures, the metabolic cost of thermoregulation is high, and the high rates of nectar consumption required to support this energy demand result in rates of water turnover that rival those of amphibians and freshwater fish³. Under these conditions, hummingbirds could reduce the metabolic cost of nitrogen excretion by producing ammonia or urea. At high temperatures, on the other hand, metabolic costs and therefore nectar consumption are lower, but water lost for evaporative cooling is high, so hummingbirds should excrete nitrogen as uric acid to reduce excretory water loss and avoid the risk of dehydration^{4,5}.

To examine nitrogen excretion in Anna's hummingbirds, with high and low rates of water turnover, we fed birds a solution of 0.7 M sucrose at three ambient temperatures. Rates of consumption at 20 °C and 40 °C were not significantly different (0.8 ± 0.1 ml h⁻¹ and 0.7 ± 0.1 ml h⁻¹, respectively; mean $\pm$ s.e.m.). At 10 °C, birds consumed 1.2 ± 0.1 ml h⁻¹, a volume of fluid that amounts to about 27% of the body mass of the bird (~4.5 g) per hour.

We collected excreted urine from birds at these three temperatures and measured the concentration of uric acid using the uricase/peroxidase method (Sigma procedure no. 685) and of ammonia and urea using the urease/Berthelot method (Sigma procedure no. 640). We were concerned about volatilization of ammonia if the urine was alkaline, but the pH of freshly excreted urine from five birds at each of the test temperatures was consistently acidic (pH 5.0–6.7).

At an ambient temperature of 40 °C hummingbirds were above their thermoneutral zone⁵, and panted after short flights and even intermittently while sitting. The resulting increase in water loss, coupled with low rates of food consumption, resulted in the excretion of small volumes of urine with concentrations of ammonia, urea and uric acid that were 3–4 times higher than those at 10 °C (Fig. 1a, c).

The concentration of ammonia in hummingbird urine was higher than that of uric acid or urea at all three temperatures (Fig. 1a, c). However, the uric acid molecule contains four nitrogen atoms whereas urea contains two and ammonia only one, thus

at both 20 °C and 40 °C ambient temperature, birds were uricotelic, with 60–65% of nitrogen excreted as uric acid (Fig. 1b). At 10 °C, five birds were ammonotelic (58% of nitrogen excreted as ammonia, 30% as uric acid; Fig. 1d). Only one bird at 10 °C was clearly uricotelic (42% of nitrogen excreted as ammonia, 51% as uric acid), whereas in the remaining four birds the amounts of nitrogen as ammonia and uric acid differed by 6% or less.

Avian uricotelic was first documented in the chicken (*Gallus domesticus*) over 200 years ago (as described in ref. 6). Since then, there have been many studies of nitrogen excretion in chickens⁷, but surprisingly few additional species have been examined — the mallard (*Anas platyrhynchos*)⁸, the pigeon (*Columba livia*)⁹, the turkey vulture (*Cathartes aura*)¹⁰, the emu (*Dromaius novaehollandiae*)¹¹ and the greylag goose (*Anser anser*)¹². These species were all found to be uricotelic, but most were studied under experimental conditions that would be expected to result in uricotelic (for example, dehydration, salt-loading, protein-loading). From these very limited data and casual observations of the appearance of a white precipitate in excreta, the ubiquity of uricotelic in birds has become widely accepted in the literature and, as far as we are aware, has never been challenged.

The ability of the Anna's hummingbird to be ammonotelic under some conditions and uricotelic under others is rare among vertebrates. Indeed, facultative ammono-uricotelic has been reported in only one other group, the crocodilians¹³. However, given the surprisingly few avian species in which nitrogen excretion has been examined, and the occurrence of atypical patterns of nitrogen excretion in other vertebrate groups (for example uricotelic in arid frogs¹⁴, ureotelic in freshwater turtles¹⁵), it is possible that non-uricotelic patterns of nitrogen excretion could occur in other birds as well.

Marion R. Preest*

Carol A. Beuchat

Department of Biology,
San Diego State University,
San Diego, California 92182, USA
mpreest@fig.cox.miami.edu

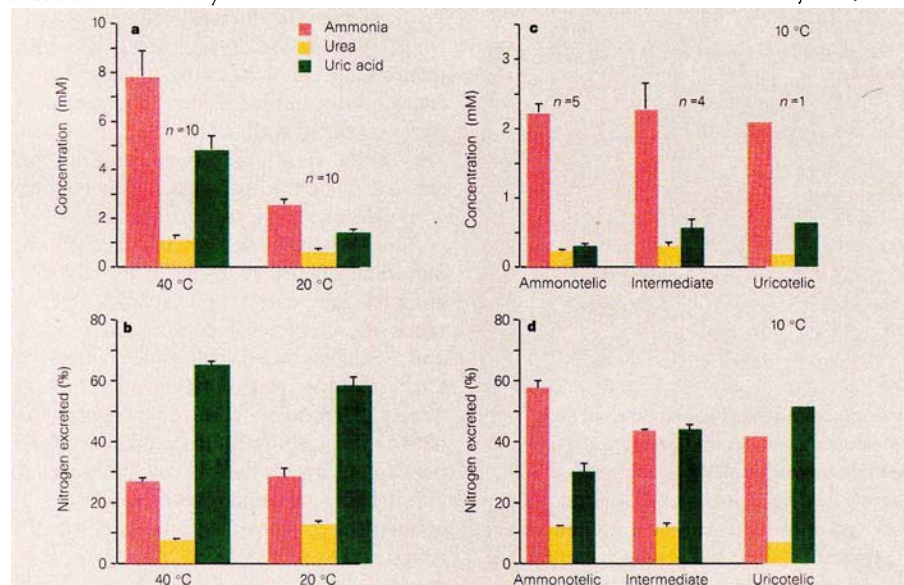


Figure 1 Nitrogen excretion of hummingbirds at ambient temperatures of 40 and 20 °C (a, b) and at 10 °C (c, d) expressed as concentration (a, c) and as percentage (b, d) of nitrogen excreted as ammonia, urea and uric acid (mean $\pm$ s.e.m.). Birds at 10 °C are separated into those that were ammonotelic, uricotelic, or intermediate (excreting roughly equal amounts of nitrogen as ammonia and uric acid).

- Campbell, J. W. in *Nitrogen Metabolism and Excretion* (eds Walsh, P. J. & Wright, P. A.) 133–161 (CRC, Boca Raton, 1995).
- Lasiewski, R. C. *Physiol. Zool.* **36**, 122–140 (1963).
- Beuchat, C. A., Calder, W. A. III & Braun, E. J. *Physiol. Zool.* **63**, 1059–1081 (1990).
- Calder, W. A. III *J. Theor. Biol.* **78**, 185–196 (1979).
- Powers, D. R. J. *Comp. Physiol. B* **162**, 74–84 (1992).
- Needham, J. *Chemical Embryology* Vol. 2 (Hafner, New York, 1931).
- Skadhauge, E. *Osmoregulation in Birds* (Springer, New York, 1981).
- Stewart, D. J., Holmes, W. N. & Fletcher, G. J. *Exp. Biol.* **50**,

9. McNabb, F. M. A., McNabb, R. A. & Ward, J. M. Jr *Comp. Biochem. Physiol.* **43**, 181-185 (1972).
10. McNabb, F. M. A., McNabb, R. A., Prather, I. D., Conner, R. N. & Adkisson, C. S. *Condor* **82**, 219-223 (1980).
11. Dawson, T. J., Maloney, S. K. & Skadhauge, E. J. *Comp. Physiol. B* **161**, 165-171 (1991).
12. Robin, J., Cherel, Y., Girard, H., Gélou, A. & Le Maho, Y. *J. Comp. Physiol. B* **157**, 491-499 (1987).
13. Coulson, R. A. & Hernandez, T. *Comp. Biochem. Physiol. B* **74**, i-182 (1983).
14. Loveridge, J. P. *Arnoldia Rhodesia* **5**, 1-6 (1970).
15. Moyle, V. *Biochem. J.* **44**, 581-584 (1949).

* Present Address: Department of Biology, University of Miami, Coral Gables, Florida 33124, USA.

Socioeconomic factors and tropical deforestation

Despite worldwide concern about habitat degradation in the tropics, deforestation in tropical countries continues unabated¹. In 1979, 75,000 km² was deforested, but by 1991 this figure had increased to 132,000 km² (ref. 2). The underlying causes of deforestation are, however, poorly understood. Here we examine the correlation between deforestation in the tropics and 14 socioeconomic variables, and show that the most significant correlates of deforestation vary from region to region. Furthermore, the relative importance of the most significant correlates of deforestation varies with the statistical model used.

Tropical forests provide important ecosystem services, and have large concentrations of biodiversity and genetic resources. Policy options concerning conservation and

Table 1 Correlation coefficients between log annual deforestation rate and selected variables

Variable (log)	All countries	Africa	Asia	Latin America
Population				
Rural population density	0.63*** (69)	0.53** (32)	0.59* (15)	0.76*** (22)
Urban population density	0.69*** (70)	0.45** (33)	0.73** (15)	0.77*** (22)
Total population density	0.68*** (68)	0.51** (32)	0.66** (15)	0.77*** (21)
Economic				
Per capita GNP	0.07 (66)	-0.16 (33)	0.23 (13)	-0.43 (20)
Per capita external debt	0.41** (62)	0.27 (33)	0.38 (11)	0.20 (18)
Forest products extracted per unit forest area				
Fuel-wood and charcoal	0.45*** (70)	-0.25 (34)	0.64* (15)	0.78*** (21)
Industrial roundwood	0.60*** (69)	0.04 (34)	0.67** (15)	0.76*** (20)
Panel-wood	0.28* (52)	-0.29 (24)	0.59* (14)	0.23 (14)
Land Use				
Cattle density	0.60*** (70)	0.46** (34)	0.09 (14)	0.88*** (22)
Crop-land area/total land area	0.74*** (69)	0.59*** (33)	0.90*** (14)	0.80*** (22)
Pasture area/total land area	0.10 (68)	-0.02 (36)	-0.12 (13)	0.79*** (19)
Land in other use/total land area	0.43*** (66)	-0.01 (33)	0.51 (13)	0.74*** (20)
Per capita energy consumption				
Traditional fuel	-0.41*** (67)	-0.17 (35)	-0.91*** (12)	-0.23 (20)
Commercial energy	0.40*** (66)	0.05 (34)	0.38 (14)	-0.13 (18)

Correlation coefficients are listed with their significance levels: *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$. The number of countries analysed is given in parentheses.

management of tropical forests are often presented in general, global terms³, reflecting a poor understanding of the causes of deforestation⁴. International comparisons of data from individual countries attribute major roles to population pressure⁵, wealth^{5,6}, external debt⁷ and competition for land^{4,8}. Deforestation is caused by many factors^{9,10}, but quantitative analysis of the relative importance of these factors throughout the tropics is lacking.

We obtained data for 70 tropical countries from the World Resources Institute database¹¹. The 14 socioeconomic variables

used were grouped into five categories: population, economic indicators, land use patterns, extraction of forest products and energy use. Pooling results over the entire tropics, deforestation rate was correlated positively with 11 of the 14 variables, and negatively with per capita traditional fuel consumption (Table 1). There were, however, strong regional differences.

For Africa, deforestation rate was more strongly correlated with rural population than with urban population, but the opposite trend was observed for Asia. Pasture-land area and the area of land in use other than as pasture or for crop growth are correlated with deforestation only in Latin America, whereas cattle density was significant in Africa and Latin America but not in Asia. With respect to forest products, panel-wood production was correlated with deforestation in Asia but not in Latin America, and none of the forest product variables were correlated with deforestation in Africa. Per capita traditional fuel consumption showed a significant negative correlation with deforestation rate in Asia alone.

Our results show that correlates of deforestation are many and varied, with complex interactions. The relative importance of each factor depends on the model and variables used. Multiple regression analysis of the relative magnitude of direct effects places proportion of crop-growing land area ahead of population density, then per capita external debt as having the greatest effect on deforestation when all tropical countries were considered, but when Latin America and Asia were looked at separately external debt was the second most important factor (Fig. 1). In Africa, population density was more important than external debt followed by crop-land area. These three variables plus fuel-wood and charcoal

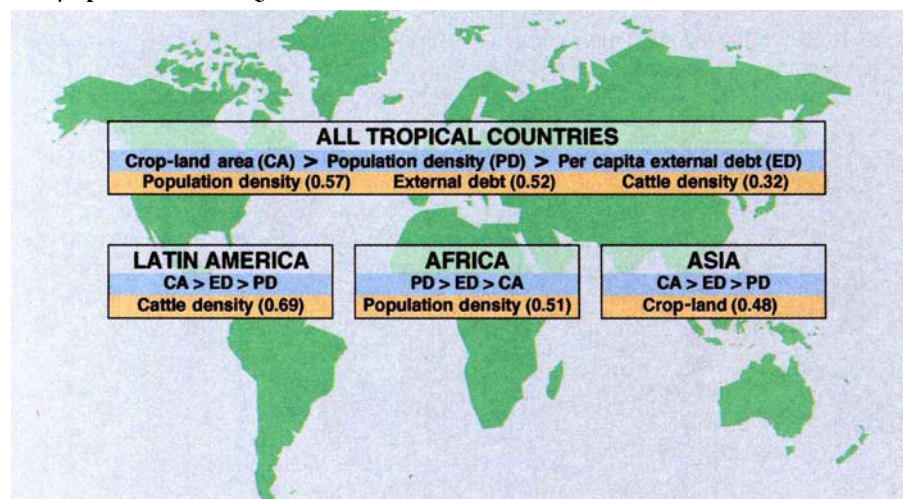


Figure 1 Relative magnitude of direct effects (β) of population density, per capita external debt, proportion of crop-land area, fuel-wood and charcoal production and per capita traditional fuel consumption on rate of deforestation as revealed by multiple regression analysis, using a model comprising one variable from each group. The variables together explained 42% (all tropical countries), 33% (Africa), 96% (Asia) and 48% (Latin America) of variance. The three most important variables are shown in blue. Results of forward stepwise regression, using independent variables that showed significant correlation ($P < 0.05$) with the rate of deforestation in tropical countries, are shown in yellow. The most significant three variables overall, and the most important in each region are shown. β is given in parentheses. Direct effects are based on standardized partial regression coefficients (β), with log annual per cent deforestation in tropical countries as the dependent variable.

production, and per capita traditional fuel consumption together explained 42% (all tropical countries), 33% (Africa), 96% (Asia) and 48% (Latin America) of variance in deforestation rate.

On the other hand, a forward stepwise regression analysis using independent variables that were significantly correlated ($P < 0.05$) with deforestation in tropical countries showed that population density ($\beta = 0.57$), external debt ($\beta = 0.52$) and cattle density ($\beta = 0.32$) are important in that order when all countries are considered together (Fig. 1). In Africa, population density has greatest effect on deforestation rate, in Asia crop-land area, and in Latin America cattle density. Again, the most significant results of our analyses are the differences between regions, also revealed by other analyses^{1,9}.

Our analyses point to several possible causes of deforestation and have important policy and management implications. The availability of remote-sensing imagery and local data on socioeconomic variables, give us the potential to identify accurately the causes of deforestation at smaller scales; as we have done for the Sarapiquí region in Costa Rica (A. Sanchez-Azofeifa, S. D. and K. S. B., unpublished results) and the western Ghats, India (S. Menon & K. S. B., unpublished results). Policies to stem deforestation must take into account regional as well as country-wide differences. More emphasis is needed on land distribution and agrarian reform in Latin America, where much of the land is used inefficiently for cattle ranching, and in Asia, alternatives must be found for subsistence-level agriculture, which drives agricultural frontiers onto marginal lands occupied by forests.

Kamaljit S. Bawa

S. Dayanandan

Department of Biology,
University of Massachusetts, Boston,
100 Morrissey Boulevard,
Boston, Massachusetts 02125, USA
e-mail: bawa@umbusky.cc.umb.edu

1. Forest Resource Assessment 1990, Survey of Tropical Forest Cover and Study of Change Processes. (FAO Forestry Pap. 130, Food and Agriculture Organization of the United Nations, Rome, 1996).
2. Myers, N. in *The Causes of Tropical Deforestation* (eds Brown, K. & Pearce, D. W.) 27–40 (University College London Press, 1994).
3. Myers, N. *Science* **268**, 823–824 (1995).
4. Brown, K. & Pearce, D. W. (eds) *The Causes of Tropical Deforestation* (University College London Press, 1994).
5. Cropper, M. & Griffiths, C. *Am. Econom. Rev.* **84**, 250–254 (1994).
6. Capistrano, A. D. & Kiker, K. B. *Ecol. Econom.* **14**, 21–29 (1995).
7. Kahn, J. R. & McDonald, J. A. *Ecol. Econom.* **12**, 107–123 (1995).
8. Pearce, D. W. & Brown, K. in *The Causes of Tropical Deforestation* (eds Brown, K. & Pearce, D. W.) 2–26 (University College London Press, 1994).
9. Rudel, T. & Roper, J. *Ambio* **25**, 160–166 (1996).
10. Kummer, D. & Sham, C. H. in *The Causes of Tropical Deforestation* (eds Brown, K. & Pearce, D. W.) 146–158 (University College London Press, 1994).
11. *World Resources 1994–1995* (World Resources Institute, Washington DC, in collaboration with the United Nations Environmental Programme and Development Programme, Oxford Univ. Press, Cambridge, 1994).

Kinked DNA

Kinks are thought to be important in DNA–protein interactions¹ but evidence for their existence in free DNA has previously been indirect. The atomic force microscope (AFM) allows direct visualization of molecular structure *in situ*^{2,3}, and using our newly developed AFM (magnetic a.c. mode or MacMode)⁴ we have observed small DNA circles change from smoothly bent to abruptly kinked shapes. This effect was dependent on the presence of specific divalent cations. The DNA was bent smoothly in the presence of Mg^{2+} , but consisted of nearly straight segments connected by kinks in the presence of Zn^{2+} .

Tandem sequence repeats of d(A)₅ and d(GGGCC[C]) bend DNA⁵. We ligated the following oligomer:

CCCCAAAAGGGCCAAAAGGGCCAAAAGGGCCAAAAGGG
TTTTTCCCGGTTTTCCTCCGGTTTTCCTCCGGTTTTCCTCCGGG

to produce DNA circles. Maximum formation of circles occurs when the ligated oligomers are 126-base-pairs long. DNA longer than this forms superhelices, and thermal fluctuations are needed to bring the ends together. As a result, the larger closed circles are axially strained on ligation. We extracted circles and purified them as described elsewhere⁶. We placed DNA solutions (0.5 mg ml⁻¹ in 1 mM Mg^{2+} or Zn^{2+}) into the sample cell of a scanning probe microscope (PicoSPM; Molecular Imaging Corp., Phoenix, Arizona) where the molecules adsorbed spontaneously onto a mica substrate. Images were obtained *in situ* and remained stable on repeated scanning.

DNA images obtained in the MacMode had an average full-width at half-height of 3.5 nm, so the instrumental broadening is little more than 1 nm, a significant improvement in resolution over previous AFM technologies. DNA circles of 168 base pairs in 1 mM $MgCl_2$ are smoothly curved, with less than one abrupt bend or kink per molecule (Fig. 1a).

When molecules from the same sample are imaged in 1 mM $ZnBr_2$, their appearance changes dramatically (Fig. 1b). The DNA consists of nearly straight segments connected by abrupt kinks. This phenomenon was not observed in 126-base-pair circles, so we conclude that axial strain is required for this conformational change. The relative lack of kinks in the presence of Mg^{2+} and the absence of kinks in smaller circles under all ionic conditions demonstrate that single-strand breaks did not cause the kinks.

The strong enhancement of kinking observed here in the presence of Zn^{2+} is qualitatively consistent with earlier electron microscopy⁷ and electro-optical⁸ studies. If

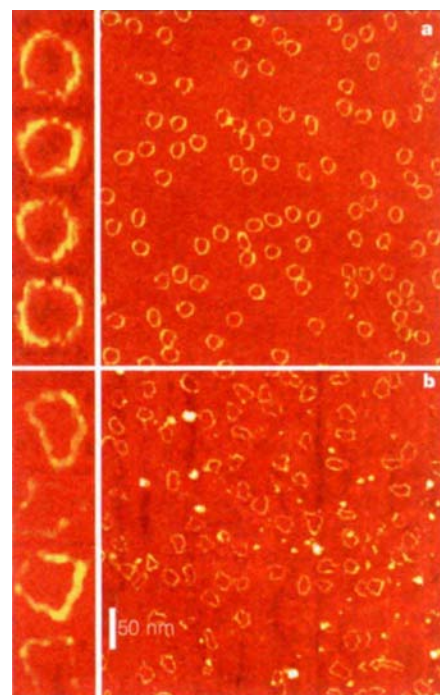


Figure 1 Images of 168-base-pair DNA minicircles in 1 mM $MgCl_2$ (a) and 1 mM $ZnBr_2$ (b), showing a fourfold increase in kink density in Zn^{2+} . Selected molecules are displayed magnified by a factor of four to the left of each image. The tip was oscillated at 25 kHz with an amplitude of 5 nm and the image was acquired in five minutes.

kinks are readily generated *in vivo* through a combination of axial strain and appropriate ionic conditions, then one might speculate that local writhing stress, to which DNA is subjected during various regulatory events, may promote the formation of localized kinks at specific DNA sequences. These could serve as signals for the assembly and/or localization of nucleoprotein complexes.

Wenhai Han

S. M. Lindsay

Department of Physics and Astronomy,
Arizona State University,
Tempe, Arizona 85287-1504, USA
e-mail: stuart.lindsay@asu.edu

Mensur Dlakic

Rodney E. Harrington

Department of Biochemistry,
University of Nevada Reno,
Reno, Nevada 89557-0014, USA

1. Werner, M. H., Gronenborn, A. M. & Clore, G. M. *Science* **271**, 778–784 (1996).
2. Hansma, H. G. & Hoh, J. *Annu. Rev. Biophys. Biomol. Struct.* **23**, 115–139 (1994).
3. Bustamante, C. & Rivetti, C. *Annu. Rev. Biophys. Biomol. Struct.* **25**, 395–429 (1996).
4. Han, W., Lindsay, S. M. & Jing, T. *Appl. Phys. Lett.* **69**, 4111–4114 (1996).
5. Dlakic, M. & Harrington, R. E. *Proc. Natl Acad. Sci. USA* **93**, 3847–3852 (1996).
6. Dlakic, M. & Harrington, R. E. *J. Biol. Chem.* **270**, 29945–29952 (1995).
7. Laundon, C. H. & Griffith, J. D. *Biochemistry* **26**, 3759–3762 (1987).
8. Nickol, J. & Rau, D. C. *J. Mol. Biol.* **228**, 1115–1123 (1992).

Biochemical typing of scrapie strains

Differences in the molecular heterogeneity of the abnormal isoforms of the prion protein, PrP^{Sc}, due to differential glycosylation and partial endogenous or exogenous proteolysis^{1,2} have been found in different scrapie strains³, sources of Creutzfeldt–Jakob disease (CJD) and fatal familial insomnia (FFI)^{4,5}. Recently Collinge *et al.*⁶ extended this analysis to include sporadic CJD and the new variant of Creutzfeldt–Jakob disease (vCJD) and found that the vCJD ‘type 4’ pattern of PrP^{Sc} was common to mice, cats and monkeys infected with bovine spongiform encephalopathy (BSE). We have extended this analysis to include scrapie strains passaged in mice and originally derived from sheep or goats.

Collinge *et al.*⁶ suggested that the prion molecular phenotype ‘type 4’ was distinct and characteristic of BSE, and that its presence in vCJD was further evidence of a direct link between vCJD and BSE. Furthermore, they suggested that this pattern, though also seen in FFI⁷, might be used to identify a BSE origin of the disease — in particular that it would differentiate BSE and natural scrapie in sheep and goats. Unfortunately, no control data of scrapie in mice or the other species were given to support this idea.

We present a glycoform ratio analysis of mouse-passaged scrapie strains from sheep or goats and compare them with a BSE-

derived strain, 301V. We confirm the original observations of Kascsak and colleagues³ that PrP^{Sc} pattern analysis can be used to differentiate strains of these agents when isolated in inbred mice. However, we have identified strains originally isolated from the same source of scrapie which have radically different patterns of PrP^{Sc}, suggesting that this may compromise its use as a tool to identify the origins of disease.

Over 15 transmissible spongiform encephalopathy strains have been experimentally derived from scrapie-infected sheep and other ruminants infected with a transmissible spongiform encephalopathy^{8,9}. They are characterized biologically by their relative incubation periods in mice of the three *Sinc* (*Prn-i*) genotypes and the distribution of vacuolation in the brain. We analysed five of these strains and a strain passaged in hamsters according to the methods of Collinge *et al.*⁶. The set was compared with 301V, derived from a BSE-infected cow by serial passage in *Sinc p7* mice¹⁰.

Each strain showed a consistent but restricted glycoform profile (Fig. 1). Individual strain glycoform profiles ranged from highly aglycosyl forms (79A) to very heavily glycosylated forms (87V and 263K). The 301V profile is similar to the ‘type 4’ pattern published for other BSE-derived sources, but is also similar to the sheep-derived 22A strain. The ME7 and 87V strains were derived from the same original source of natural scrapie¹¹ and yet they show very different glycoform profiles. Similarly 79A (mouse) and 263K (hamster) derive from the same source¹² but show extreme differences in glycoform profile, albeit in different species.

Our results show that, under controlled experimental conditions, the glycoform ratio can be used to differentiate between strains of transmissible spongiform encephalopathies, but, like other strain properties, it may change at the primary or subsequent passage and not necessarily reflect the characteristics of its progenitor strain^{13,14}. Much more information is needed about the contributions of source of agent and host PrP genotype to the glycoform ratio before it can be used as a valid predictor of the origin of a particular transmissible spongiform encephalopathy.

**Robert A. Somerville, Angela Chong
Orla U. Mulqueen**

*BBSRC & MRC Neuropathogenesis Unit,
Institute for Animal Health, West Mains Road,
Edinburgh EH9 3JF, UK*

Christopher R. Birkett

Sophie C. E. R. Wood, James Hope

*Institute for Animal Health, Compton,
Newbury, Berkshire RG20 7NN, UK
e-mail: james.hope@bbsrc.ac.uk*

Collinge et al. reply — We recently reported that new variant CJD (vCJD) could be distinguished from previously recognized forms of sporadic and acquired CJD by a consistent

pattern on western blots with respect to ratios of PrP^{Sc} glycoforms after proteinase K digestion⁶. This pattern resembled that seen in BSE in cattle and BSE transmitted to mouse, domestic cat and macaque, consistent with the hypothesis that BSE causes vCJD. We did not analyse all prion strains in all species, so the type of pattern that we saw in the ratio of PrP^{Sc} glycoforms may not be unique to BSE and vCJD. We suggested that PrP^{Sc} analysis might be helpful in differentiating BSE and scrapie in sheep to encourage molecular studies in addition to conventional approaches. This approach may be complicated, however, by the multiplicity of PrP alleles in sheep, the many different strains that have been isolated from natural sheep scrapie, and the uncertainty as to whether BSE originated in sheep.

Somerville *et al.* subjected scrapie strains that had been passaged in mice to PrP^{Sc} glycoform analysis. They show that several of these ‘classical’ scrapie strains, previously well characterized by conventional transmission studies and lesion profiling in inbred mice, can also be differentiated by this technique, providing further support for the use of molecular analysis of prion strains.

It is unsurprising that not all of these strains can be completely differentiated from BSE. Given the spread of the clusters on our glycoform scattergrams, it would have been remarkable if none of these scrapie strains showed any overlap with BSE or CJD. This type of analysis remains relatively crude; higher-resolution molecular techniques are needed. But despite the limitations and noise inherent in the current methods, several classical scrapie strains can now be differentiated from BSE in mice, in days rather than years.

John Collinge

Andrew F. Hill, Katie C. L. Sidle

*Neurogenetics Unit,
Imperial College School of Medicine at
St Mary's Hospital,
London W2 1PG, UK*

James Ironside

*National CJD Surveillance Unit,
Western General Hospital, Edinburgh EH4 2XU, UK*

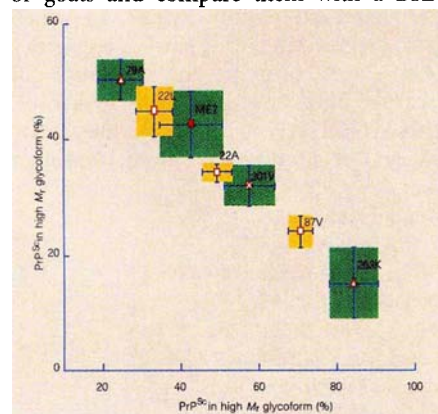


Figure 1 Glycoform profiles were obtained from PrP^{Sc} extracted from whole brains of rodents infected with scrapie or BSE passaged in mice, using an antipeptide rabbit antibody (R30, PrP 90–104)¹⁵ donated by B. Caughey. The models used included 79A and ME7 scrapie strains in the SV mouse strain (*Sinc s7*), 22L in C3H mice (*Sinc s7*), 22A, 87V and 301V in VM mice (*Sinc p7*) and 263K in Syrian hamsters. The glycoform ratio of a particular strain of scrapie is independent of *Sinc* (*Prn-i*) (results not shown). The scattergram shows the percentage of PrP^{Sc} in the high-molecular-mass glycoform plotted against the percentage in the low-molecular-mass glycoform. Error bars, s.d.; *n* = 3–9.

- Hope, J. *et al.* *Eur. J. Biochem.* **172**, 271–277 (1988).
- Somerville, R. A. & Ritchie, L. A. *J. Gen. Virol.* **71**, 833–839 (1990).
- Kascsak, R. J. *et al.* *J. Gen. Virol.* **66**, 1715–1722 (1985).
- Parchi, P. *et al.* *J. Neuropathol. Exp. Neurol.* **54**, 416 (1995).
- Gambetti, P. in *Prions, Prions* (ed. Prusiner, S. B.) 19–25 (Springer, Berlin, 1996).
- Collinge, J., Sidle, K. C. L., Meads, J., Ironside, J. & Hill, A. F. *Nature* **383**, 685–690 (1996).
- Telling, G. C. *et al.* *Science* **274**, 2079–2082 (1996).
- Bruce, M. E., McConnell, I., Fraser, H. & Dickinson, A. G. *J. Gen. Virol.* **72**, 595–603 (1991).
- Bruce, M. *Prion Dis.* **13**, 223–236 (1996).
- Bruce, M. *et al.* *Phil. Trans. R. Soc. Lond. B* **343**, 405–411 (1994).
- Bruce, M. E. & Dickinson, A. G. in *Slow Transmissible Diseases of the Central Nervous System* (eds Prusiner, S. B. & Hadlow, W. I.) 71–86 (Academic, New York, 1979).
- Kimberlin, R. H., Walker, C. A. & Fraser, H. *J. Gen. Virol.* **70**, 2017–2025 (1989).
- Bruce, M. E. & Dickinson, A. G. *J. Gen. Virol.* **68**, 79–89 (1987).
- Lazmezas, C. I. *et al.* *Science* **275**, 402–405 (1997).
- Kocisko, D. A., Lansbury, P. T. & Caughey, B. *Biochemistry* **35**, 13434–13442 (1996).

Journals of our plague years

And the Waters Turned to Blood

by Rodney Barker

Simon & Schuster: 1997. Pp. 334. \$24

Deadly Feasts: Tracking the Secrets of a Terrifying New Plague

by Richard Rhodes

Simon & Schuster: 1997. Pp. 250. \$24, £14.99

Robert Desowitz

If Daniel Defoe were alive today, he would be known as a science writer on the strength of his *A Journal of the Plague Year* (1722). Richard Rhodes and Rodney Barker's books are journals of our own plague years. The authors, both of whom are journalists, differ in style, comprehensibility and calibre of writing; and their 'diseases' are very, very different. But they both present their accounts in a dramatic, near-novel-like fashion, each with a clearly identifiable hero or heroine (each somewhat humanly flawed), sub-heroes, and hissable villains in the guise of nasty scientists and even nastier government agencies heeding only special interests.

The setting of Barker's book is North Carolina, a state that cherishes the natural beauty of its shores, streams, rivers and mountains. I write this from my deck overlooking a small lake in the longleaf piney Sandhills. The trees are beginning their spring flowering spectacular and a red cockaded woodpecker is rapping away. It is hard to imagine that so pristine a place could become blighted. But in 1990, many millions of fish died in the Neuse and Pamlico rivers and their estuaries.

About the same time, fish began dying in the laboratory tanks of Edward Noga, a fish pathologist at North Carolina State University School of Veterinary Medicine. He calls on JoAnn Bukholder (the heroine), a newly appointed assistant professor of phycology at the university, for advice. Although an algae specialist, she identifies a dinoflagellate of the genus *Pfiesteria* as the pathogen in the rivers and Noga's tanks. This is a microorganism of fresh and brackish water previously unrecognized for its pathogenicity. So why did *Pfiesteria* bloom in 1990?

Bukholder traces the cause to an egregious violation of the environment. The state government has pandered to the politically powerful hog farmers and allowed them to dump millions of gallons of pig faeces into the rivers — to the biological delight of the dinoflagellate. She begins a crusade to halt the pollution and to cleanse the rivers. She fights for grants to continue her experimental and ecological studies. She fights the secretive Noga. She fights the chief and staff of the Division of Environmental Management. She accuses the governor of putting pigs before people. And then comes the unsettling evidence that not only are the rivers and fish

sick because of the dinoflagellate, but people are too.

Bukholder develops a fever; there is a continuing lassitude and memory disturbance like that of an early Alzheimer's victim. Her assistants come down with similar neurological manifestations. They leave the laboratory and slowly recover. Fishermen of the rivers and estuaries have a similar syndrome. Bukholder now believes that the dinoflagellate is toxic to humans. The Department of Health is reluctant to accept the thesis without further proof. But that proof, from this account at least, has not been forthcoming. As with 'Gulf War syndrome', Koch's postulates remain unsatisfied.

The story ends rather inconclusively. Barker tells a good story, but descriptive science is not his strength. This is in contrast to Rhodes, who knows his way around science and scientists and whose *Deadly Feasts* is a scholarly, fascinating account of the transmissible spongiform encephalopathies.

Humans and their domestic animals are being killed by what appears to be a new entity. It is transmissible like a microbe but gives no signs of life, no morphological life structure, and has no obedience to life's canon of DNA to RNA to protein. It has no discernible nucleic acid and excites no immune response. All who are stricken die. The diseases are associated with amyloid fibrils that have been called prions (as well as virinos, slow viruses and amyloid rods). The latest, but undoubtedly not the last, word is that they are amyloid fibres polymerized from abnormal monomeric proteins by mutations in lysozyme (Booth *et al.*, *Nature* 385, 787–793; 1997).

In the 1950s it came to the attention of the Australian colonial authorities that unusual deaths were occurring among the Fore tribe in the eastern highlands of Papua New Guinea. The natives called it kuru, the shivering death. Only the women and children died of this strange progressive neurological disease from which there was no recovery. The natives knew it as witchcraft. The Australians did not know what to make of it; no aetiological agent could be found.

Into this confused scene came, in 1957, Carelton Gajdusek, the American paediatrician, virologist, anthropologist and non-stop monologist. It was his guiding genius that solved the mystery of kuru and the related 'slow virus' diseases — notably, Creutzfeldt-Jakob disease and bovine spongiform encephalopathy (BSE), 'mad cow disease'.

Gajdusek and his Australian colleagues, Michael Alpers and Vincent Zigas, showed that cannibalism was the cause of kuru. The Fore women and children, but not the men, ate — totally consumed — their dead relatives.

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

A patient with kuru, the 'shivering death'.

Kuru brains were sent to Gajdusek's laboratory at the US National Institutes of Health in Bethesda, Maryland. Chimps were inoculated and after a long period they came down with the typical symptoms and died. Kuru was shown to be a transmissible disease but no infectious agent was detected.

William J. Hadlow, the American research veterinarian, brought Gajdusek's attention to the similarities between scrapie, the disease of sheep he had been studying, and kuru. The penny dropped that this is a whole new galaxy of disease. Stanley Prusiner, whom Rhodes does not like very much, bestowed the name prions on the amyloid rods, but Gajdusek won the Nobel prize.

In the last chapters, Rhodes describes the newest prion disease, BSE. He reserves his most scathing anger and contempt for the British health authorities, which refused for so long to recognize the threat when it became apparent that humans were susceptible. The animal feed industry is castigated for faulty processing of animal parts. The Swiss, Rhodes reports, were selling human placentas from abortion clinics to animal feed processors. Incredible! Rhodes serves as a kind of latter-day Sinclair Lewis, who excoriated the Chicago meat-packing industry.

The last words of Rhodes come from Gajdusek, who pronounces all amyloid diseases to be essentially the same and warns that they could bring down the human species.

This reviewer's last words are for Gajdusek, for whom this is a troubled time, to remind him that all humanity is indebted to his genius. □

Robert Desowitz is in the Department of Epidemiology, School of Public Health, University of North Carolina, Chapel Hill, North Carolina 27599, USA.

Teachers should go back to school

Talking About Leaving: Why Undergraduates Leave the Sciences

by Elaine Seymour and Nancy M. Hewitt
Westview Press: 1996. Pp. 429. \$49.95, £37

Cynthia M. Friend

Elaine Seymour and Nancy Hewitt seek to understand why US undergraduate students leave science and engineering majors, and to identify areas where intervention could have a positive effect. The motivation for their study was derived from the widely publicized belief that the fundamental approach to science and mathematics education was flawed, leading to a waste of talent.

The impetus for US science education reform is rooted in several reports issued in the late 1980s and early 1990s that sound the alarm by predicting a shortage of scientists and engineers in the next century. Recruitment and retention of women and minorities to the sciences was generally offered as part of the solution to the under-supply problem. Although recent under-employment of scientists, due to changes in

the political and economic landscape, demonstrates that a shortage is not imminent, improving science teaching remains the focus of many organizations and scientists. Educational reform in the sciences and mathematics is currently motivated by a perceived need for science literacy and for increased diversity in technical disciplines.

The first chapter of *Talking about Leaving* provides a thorough outline of the recent history leading up to the authors' study, making this an excellent reference source for those interested in issues of science and education reform.

Seymour and Hewitt's conclusions are thought-provoking and somewhat surprising, and will cause science and mathematics educators to re-examine their assumptions about why students do and don't become scientists. The authors conclude that switching out of a science and engineering major is not due to students discovering that they erred in selecting their major, nor is it directly related to large class size. Furthermore, the grades earned by "switchers" and "non-switchers", using the parlance of the book, are not substantially different.

Various factors — such as lack of interest in science, interest in other disciplines, poor teaching, the workload and pace of science courses, financial factors, the level of high-

school preparation, advising, and conceptual difficulties — are discussed in detail and their relative importance weighed, on the basis of the responses of undergraduate students at several different types of institutions.

The quality of teaching is identified as a key factor in the loss of talent, across gender and ethnic lines. The debilitating effect of the old-style "weed-out" system, whereby introductory courses are made into a test of fire, is correctly identified as antithetical to the learning process and a source of talent "wastage". If nothing else, the old-style pedagogy, which requires students to devote themselves to science and places undue emphasis on exams, serves as a barrier to developing science literacy in students who ultimately choose another career path. That is a particular problem when one considers the importance of science literacy for law-makers, business people and citizens in our increasingly technological society. The "weed-out" system is particularly damaging in terms of gender and ethnic diversity, because it eliminates individuals who may have had less rigorous training in science and mathematics in high school.

Issues of special importance in the retention of women and underrepresented minorities are also thoughtfully considered. There is clear evidence that female students and those from minorities will benefit from improvements in science teaching. Indeed, peer attitudes and teaching are identified as central factors in retaining female students. The strong point of the chapter on race and ethnicity is the recognition that a diverse range of ethnicity and culture is often treated as the same, even though different groups might have a specific need or barrier to a career in science or engineering.

Although overall the methodology of Seymour and Hewitt is sound, I found the assertions of an inherent conflict between teaching and research activities to be oversimplified and not well documented. Seymour and Hewitt tacitly accepted students' perceptions that the research activities of a faculty member drive a disinterest in teaching, without examining the facts or providing corroborating evidence through interviews with the faculty members who were teaching the students.

I suspect that the perceptions of students were coloured by their expectations of what constitutes good science teaching based on their experiences in secondary school, where emphasis is placed on rote rather than conceptual learning. Indeed, there is a contradiction in their thesis since students expressed a desire to have current and relevant information included in course material. Who best to incorporate the latest results if not someone actively engaged in research? It is often the case that



Film star walks on water

The world of the insect was the subject of the Miramax film *Microcosmos*, an official selection at the 1996 Cannes Film Festival. To accompany the film, Claude Nauridsany and Marie

Pérennou have brought the stars (in this case, just to be contrary, a pirate spider) sharply into focus in *Microcosmos: The Invisible World of Insects* (Stewart, Tabin & Chang, \$35, £23).

individuals who are excited about their research are also engaging and dynamic teachers who incorporate innovative tools into their courses. I was disappointed that Seymour and Hewitt did not examine the underlying reasons for the students' perceptions and the accuracy of the students' criticism.

Talking about Leaving does not make easy reading — I do not recommend it for reading lying down or at the beach. The density of information is extremely high and I found that I needed to take notes and highlight key points in the text to be able to extract the essence of the conclusions. The text is also rife with quotations from students interviewed for the study which were intended to exemplify the point. The effect is to distract the reader from the train of thought in the surrounding text, and I found myself needing to re-read previous material to develop a conceptual thread.

Nevertheless, this is a good reference text for those interested in improving undergraduate curricula in science and engineering. The many quotations and anecdotes will be useful to anyone who needs to illustrate problems in science and engineering education. □

Cynthia M. Friend is in the Department of Chemistry, Harvard University, Cambridge, Massachusetts 02138, USA.

You can see right through me

Naked to the Bone: Medical Imaging in the Twentieth Century

by Bettyann Holtzmann Kevles
Rutgers University Press: 1997. Pp. 378.
\$35.95

John Mallard

The ability to see through our bodies has shown us rather more these past one hundred years or so since X-rays and radioactivity were discovered than we may have realized. Until I read this book by Bettyann Kevles I had no idea that the discovery of X-rays was the inspiration for H. G. Wells to write *The Invisible Man*, or that Maurice Ravel and George Gershwin had brain abnormalities, which were revealed by the use of X-ray contrast agents.

The book is part of a series that aims to convey "the technical and human dimensions of the subject; the invention and effort entailed in devising the technologies; and the comforts and the stresses they have introduced into contemporary life". These goals are achieved very well, as you might expect in a book by a journalist with a newspaper science column who has won an award for the best science book for young adults and teaches science to art students.

Umpteenth editions and other revisions

Behavioural Ecology: An Evolutionary Approach. Fourth edition

by J.R. Krebs and N.B. Davies
Blackwell Science, £26.50, \$49.95

Like the previous three editions of this classic behavioural ecology text, this fourth edition brings together a completely new set of chapters to provide an up-to-date picture of the subject. When T.R. Birkhead reviewed the second edition in *Nature* 314, 37 (1985), he wrote: "All in all this is an excellent text. It is well produced, clearly written and provides plenty to think about". This latest edition maintains these high standards.

The Evolution of the Soul. Revised edition

by Richard Swinburne
Oxford University Press, £40, \$75 (hbk); £15.99, \$24.99 (pbk)

Richard Swinburne argues that we can only make sense of the interdependence of mental states and brain states by supposing that "mental states are states of a soul, a mental substance in interaction with the body". When D.M. MacKay reviewed the first edition in *Nature* 323, 679 (1986), he wrote: "I do not know of a better or more persuasive statement of the case for dualist interactionism in the sense he defines".

Animal Physiology: Adaptation and Environment. Fifth edition

by Knut Schmidt-Nielsen
Cambridge University Press, £24.95, \$47.95
This highly acclaimed undergraduate textbook for courses in physiology and zoology was

first published in 1975. This fifth edition takes into account recent developments in the field, particularly in movement and biomechanics.

Optima for Animals. Revised edition

by R. McNeill Alexander
Princeton University Press, \$60, £45 (hbk); \$24.95, £18.95 (pbk). Distributed in the UK by Wiley
The optimizing processes of evolution and learning have shaped the structures, movements, behaviour and life histories of animals. Here they are explained for the undergraduate, with the mathematics kept as simple as possible.

Biogeochemistry: An Analysis of Global Change. Second edition

by William H. Schlesinger
Academic Press, \$49.95, £34.95
This undergraduate text covers the effects of life on the chemistry of the Earth. It has been updated to take into account the rapid changes in our understanding of the way the Earth's systems function.

A Dictionary of Genetics. Fifth edition

by Robert C. King and William D. Stansfield
Oxford University Press, £35, \$45 (hbk); £19.95, \$24.95 (pbk)
From A to zymogen granules, the fifth edition of this dictionary contains over 6,500 definitions relevant to geneticists. It also has a classification of living organisms, a section on domesticated species, and a chronology of over 400 years of genetic research.

There are very good numbered notes for each chapter, and a most interesting bibliography separated into history, culture and society; law and policy; fiction; and medicine, science and technology.

The first part of the book, dealing with the first half-century of X-rays, is the most reliable. However, in the introduction, Wilhelm Röntgen is described as "a somewhat pedestrian German physicist" — this is later dispelled when "Röntgen's genius" is explained, and I should think so, too! Although it is brought out that he did not patent his discovery, it is not mentioned that he died a pauper. Philipp Lenard's jealousy is well described, the atmosphere of the times is captured, and the early days of X-ray tomography, before computed tomography, are well discussed.

The author seems to have lost her way a bit in the second half, which deals with the newer imaging methods of computed tomography, magnetic resonance imaging, positron emission tomography and ultrasound, in that order. It is out of sequence, however, as nuclear medicine imaging, including that with computed tomography, came before X-ray computed

tomography, and magnetic resonance imaging came after nuclear medicine imaging and positron emission tomography. The "Timeline" table at the end of the book helps appreciation of the timeflow and the parallel developments in the various

Hand in marriage: Röntgen used his wife's hand for this first X-ray image of a human.

modes of imaging.

The early days of the evolution of nuclear medicine imaging are completely missed out, and the book says that the first computerized image was an X-ray computed tomography one, but it was actually from a gamma camera. The social side is well covered, but is a bit obsessed with violence and the assassinations of US presidents, as these are a means of illustrating the use, or not, of the appropriate imaging mode of the time.

The book is clearly intended for readers in the United States. Whether the US chauvinism is deliberate or not, the effect is the same: the playing down of UK and European contributions. *Nature* urges book reviewers to avoid parochialisms as they can be irritating to readers in other countries; I found the US parochialisms in this book downright annoying. The British contributions to nuclear medicine, positron emission tomography, ultrasound and particularly to magnetic resonance imaging are largely ignored. Of the 46 people Kevles interviewed for the book, only two were UK scientists. The index references 21 US institutions but only two British and one German. The first radiological society in the world was founded in 1897 in Britain, but is

not even mentioned. It seems that history is being rewritten by journalists; perhaps history is always like this. Can one really rely on what it says in the history books? Once one author gets it wrong, then it may become perpetuated in books that follow. There ought to be more history books written by the scientists themselves.

Examples of this chauvinism are legion. In the chapter on ultrasound I could find no mention of the Nuclear Enterprises Diasonograph, which was one of the leading commercial instruments for many years; Ian Donald's pioneering work in Glasgow is not mentioned until right at the end, and then somewhat grudgingly; it says that in 1980 magnetic resonance imaging was not clinically useful, whereas British National Health Service patients in Aberdeen were being usefully imaged from August 1980 onwards.

I must declare an interest here as I have devoted much of my working life to developing medical imaging, particularly in nuclear medicine and magnetic resonance imaging. I therefore find it rather galling that most of it was completely ignored; that, when I was mentioned, I was described as a medical physician, when I am, in fact, a physicist; and that my major contribution to magnetic res-

onance imaging — the team breakthrough that made it useful clinically — was credited to others in the United States! Neither is credit given to Ian Young's team in London for making the first superconducting magnet imager work at Hammersmith Hospital in 1981. There is also no mention of Japanese work until the chapter on ultrasound, and no mention that for years Japan was second only to the United States in its take-up of magnetic resonance imaging. Perhaps this book's subtitle should have been "Medical Imaging in the United States in the Twentieth Century".

The final chapter, entitled "The Transparent Body in Late Twentieth Century Culture", is fascinating, though. To quote: "The reality of seeing through and into the living body has continued to seep into cultural consciousness, transforming our sense of what can be and ought to be seen". This book may help people to appreciate and respect science by helping to show its worldwide influence on our culture and thinking. □

John Mallard is professor emeritus of the Department of Bio-medical Physics and Bio-engineering, University of Aberdeen, Foresterhill, Aberdeen AB9 2ZD, UK.

Specialist books

Excellent ★★★★★ Good ★★★★ Average ★★★ Poor ★

Early Vertebrates

by Philippe Janvier

Oxford University Press: 1996. Pp. 393. £75, \$135

Philippe Janvier has excelled himself with this landmark publication, which surely ranks as a modern classic. No other textbook has dealt so clearly with the origins and early evolution of vertebrates. Each chapter is lucidly written and important anatomical features are clearly illustrated, providing just the right level of technical information without getting bogged down with details. Competing evolutionary relationships are cleverly depicted on opposite sides of certain illustrations, driving home the message that many current issues are unresolved. This is an essential source book for teachers and students of vertebrate anatomy and evolution.

Anatomy lies at the core of the book, but there are also useful chapters on Recent vertebrate diversity, phylogenetic principles, the fossil record, and current theories and controversies. There are also entertaining chapters on the history of important early vertebrate discoveries and discoverers, from whom I note Janvier modestly excluded himself; that was the most serious omission I could find.

John G. Maisey Department of Vertebrate Paleontology, American Museum of Natural History, New York, USA.

Range	★★★★★
Depth	★★★★★
Accuracy	★★★★★
Up-to-dateness	★★★★★
Accessibility	★★★★★
Style	★★★★★

Supersymmetry in Disorder and Chaos

by Konstantin Efetov

Cambridge University Press: 1997. Pp. 441. £65, \$100

Those strangely attracted to chaotic dynamics by the fearful symmetries of nature need read no further. Supersymmetry is a serious business, having little in common with the phenomenal fantasies that share the name.

This super-saturated manifesto from the Landau Institute is on the spectre of supermathematics of commuting and anticommuting Grassmann variables. Ground rules are provided for a game, a method of calculation that leads from random field theory to quantum chaos, for those proficient in mathematical analysis, complex variables and quantum mechanics. The rules are applicable to the theoretical, and not just imaginary, physics of disordered systems of practical importance, leading to localization and mesoscopic and quantum effects. Its power over perturbative approaches is illustrated by applications to disordered metals and semiconductors, including mesoscopic quantum devices that are emerging as the basis for nanotechnologies.

Arun Holden Centre for Nonlinear Studies, University of Leeds, UK.

Range	★★★★★
Depth	★★★★★
Accuracy	★★★★★
Up-to-dateness	★★★★
Accessibility	★★★
Style	★★★

Concise Thermodynamics

by Jeremy Dunning-Davies

Albion: 1996. Pp. 126. £15

Thermodynamics is such a widely used discipline that any new book on the subject warrants attention. This pleasant and slim volume offers a rapid and simple introduction that can be used by physicists, chemists and engineers.

However, one would not expect such a small book to introduce reaction rates (useful for chemists), actual efficiencies of power plants (useful for engineers) or the entropy of mixing (useful for solid-state physicists and chemists). You get what you pay for, namely the bare bones of thermodynamics. There are problems and solutions that experts will have seen before but that will help the beginner to sort out his ideas.

For those wanting just a skeleton introduction, this book may well serve their needs. The last chapter on black holes and the concluding remarks about possible thermodynamic errors in cosmology, however, will not be acceptable to the physics community, and would be better omitted in any future edition.

Peter T. Landsberg Faculty of Mathematical Studies, University of Southampton, UK.

Range	★★★
Depth	★★★
Accuracy	★★★
Up-to-dateness	★★★
Accessibility	★★★
Style	★★★

Transcriptional activation by recruitment

Mark Ptashne & Alexander Gann

The recruitment model for gene activation stipulates that an activator works by bringing the transcriptional machinery to the DNA. Recent experiments in bacteria and yeast indicate that many genes can be activated by this mechanism. These findings have implications for our understanding of the nature of activating regions and their targets, and for the role of histones in gene regulation.

For some years now we have understood, in broad outline, the principles of gene regulation in eukaryotes. That picture might be sketched as follows. Eukaryotic transcriptional activators stimulate transcription of otherwise silent genes. The typical activator, like its bacterial counterpart, activates by binding to specific sites on DNA and, with its 'activating region', contacting the multiprotein machinery that directs transcription (Fig. 1). For many activators and genes the specificity of activation is determined solely by the DNA-binding address of the activator. For example, the activator Gal4 ordinarily activates genes required for galactose metabolism in yeast. But when any of a wide array of genes is modified so as to bear Gal4 binding sites nearby, Gal4 activates that gene as well. Gal4 will also work in higher eukaryotes, and, as in yeast, can do so cooperatively with other DNA-binding activators. These observations suggest that the underlying mechanisms of activation are similar in different eukaryotes, and that regulatory networks—which include specific DNA-bound repressors that counter the effects of activators—can evolve by distributing binding sites for regulatory proteins near the genes that are to be regulated (for a review, see ref. 1).

Further illustrations and elaborations of this scheme have emerged over the past few years, but still unresolved are an array of questions concerning the nature of activating regions and how they work. For example, how do activating regions—families of peptides with little obvious similarities, even fragments of which still retain function—exert their effects on their target proteins?

What are those targets? Are specialized co-activator proteins required to link activating regions to the transcriptional machinery? What complications are introduced by the fact that the typical gene is wrapped on histones to form nucleosomes, which present a formidable barrier to the transcriptional machinery?

We begin by summarizing recent findings in bacteria that characterize a basic mechanism for activation called recruitment. We describe how a large variety of activator–target interactions can effect gene activation by this mechanism. We then describe yeast experiments indicating that a similar mechanism works in that eukaryote. The model that emerges at least partly explains activating region design, the nature of the targets of those peptides, and how disparate activators can work cooperatively. The model also suggests a role for histones in the process.

Bacteria

The machinery

All genes in *Escherichia coli* are transcribed by an RNA polymerase with four essential subunits (Fig. 2). Most genes are transcribed by an enzyme bearing the sigma 70 (σ^{70}) subunit, and certain other genes, an example of which we will encounter, are transcribed by a form of the enzyme bearing a different σ subunit. The polymerase recognizes characteristic promoter sequences found upstream of the transcription start site. Certain promoter sequences direct efficient transcription initiation in the absence of activators, whereas others require assistance from activators. A considerable

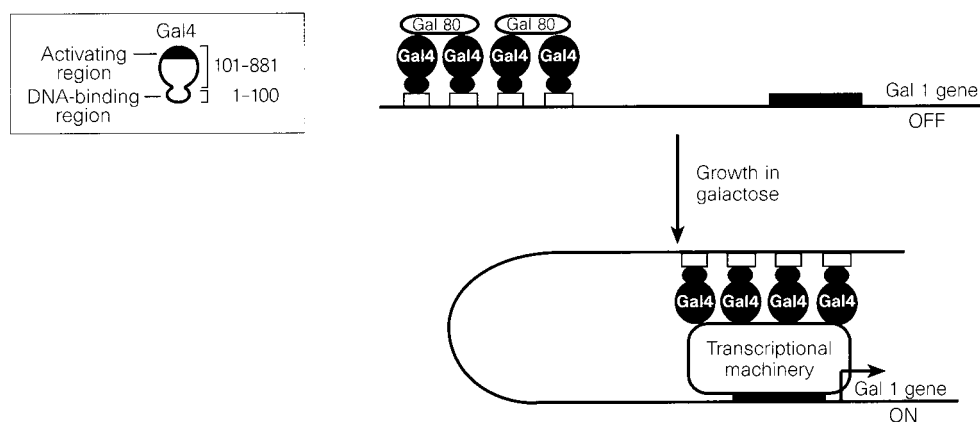


Figure 1 The yeast activator Gal4 binds to sites located approximately 250 base pairs upstream of *GAL1*, one of the genes required for galactose metabolism in yeast. Similar Gal4 binding sites are found upstream of other *GAL* genes in yeast. In the absence of galactose from the growth medium, the activating regions of Gal4 are covered by the inhibitor Gal80. Growth in galactose releases that inhibition, and Gal4's activating region contacts the transcriptional machinery to

trigger activation. The DNA between the activators and the gene loops out to accommodate the reaction. The representation is highly schematic; there are actually four Gal4 sites upstream of *GAL1*, each of which is recognized by a Gal4 dimer. The DNA-binding and activating regions are found on separable domains, as indicated.

BOX Revisiting activation in bacteria

In the text we draw parallels between the action of typical yeast activators and that of the bacterial activators CAP and lambda repressor, and we contrast this shared mechanism with that of another bacterial activator, NTRC. We say that the first group of bacterial activators (CAP and lambda repressor) recruit polymerase to DNA, whereas NTRC works on a stably bound polymerase. This distinction, the significance of which we expand upon here, partly reflects differences between the promoters at which these activators work. At all bacterial promoters, polymerase binds in at least two steps: in the first, it binds to helical DNA to form a 'closed' complex, and in the second, the complex isomerizes to form the 'open' complex in which the DNA strands are locally opened⁸⁸. But there is a crucial difference in the stabilities of the closed complexes, and in the requirements for the transition from the closed to the open complexes, at these two classes of promoters.

Consider first the typical promoter (for example, that of gene *glnAp2*) stimulated by NTRC. Experiments *in vitro* show that polymerase (bearing σ^{54}) forms a highly stable closed complex with DNA in the absence of activator, and *in vivo* polymerase is found bound to the promoter before the action of activator^{89,90}. At that promoter, therefore, the activator's role is not to stimulate formation of a stable polymerase-DNA complex (that is, to recruit polymerase), but rather to stimulate isomerization from one stable complex (closed) to another (open). In this case, therefore, isomerization, a reaction that occurs extremely rarely in the absence of activator at this promoter, is a post-recruitment step. Consistent with the idea that the activator works on a prebound polymerase, NTRC (and its relative NifA), expressed at sufficiently high levels, will activate its target genes even if it (the activator) cannot bind DNA^{31,32}.

Now consider, in contrast, promoters stimulated by CAP and lambda repressor. At these promoters, polymerase (bearing σ^{70}) is not stably associated before activation⁹¹⁻⁹³ and so we say that these activators work by recruitment. The closed complex is less stable than that formed at *glnAp2*, but once formed, the transition to the open complex can occur more readily than at the *glnAp2* promoter. Because polymerase is bound stably only upon formation of the open complex, in these cases isomerization is part of the recruitment process. The activators recruit by stimulating one or another, or both, of the steps that lead to formation of the stable open complex. For example, CAP working at the *lac* promoter primarily stimulates the initial binding of polymerase⁹⁴, lambda repressor primarily stimulates isomerization at its own promoter (called P_{RM})⁹⁵, and CAP stimulates both steps at the *gal* promoter¹¹. The stimulation of either or both steps has the effect of moving the promoter from one state, in which polymerase is largely absent, to another, in which polymerase is bound and can initiate transcription.

Strikingly, despite these apparent differences in activation at the *lac*

promoter and at P_{RM} , artificial tethering of polymerase to DNA results in significant activation at both promoters (Fig. 4). It is obvious how this might work at the *lac* promoter: in that case the artificial tethering would mimic the effect of CAP in holding the polymerase on the DNA, where it spontaneously isomerizes. The explanation for the effect at P_{RM} is less obvious, but the phenomenon can readily be accounted for in our scheme as follows. Polymerase, which binds only intermittently to this promoter in the absence of activator, typically dissociates before isomerizing. DNA-bound lambda repressor stimulates the isomerization step, and thus activated levels of transcription are achieved by ensuring that even a fleetingly bound polymerase gets pushed into the stable open complex. A plausible explanation for activation by artificial tethering at P_{RM} , therefore, would be that the tethering increases the fraction of time the promoter is occupied by polymerase and thereby increases the likelihood that isomerization would occur. A similar explanation—increased formation of the closed complex—would explain the observation⁹⁶ that, *in vitro* high concentrations of polymerase obviate the need for stimulation by repressor for activated levels of transcription⁹⁶, and the finding that, *in vivo*, a suitably positioned CAP molecule can activate P_{RM} , presumably by stimulating formation of the closed complex²².

We have noted that, in recruiting polymerase to a stable complex with DNA, lambda repressor and CAP stimulate different steps in the reaction at P_{RM} and the *lac* promoter, respectively. These steps leading to recruitment are analogous to steps in a typical enzymatic reaction (K. Ebright, personal communication) an analogy suggesting that similar protein-protein interactions can account for the action of both activators. Thus, in the enzyme-substrate example, the same interactions between enzyme and modified forms of substrate can facilitate different steps of the reaction (K_m , K_{cat} , or both). By analogy, the step in initiation affected by a given activator-polymerase-promoter interaction would depend on the positioning of the components, the identity of the promoter sequence, and the properties of the polymerase. A recent finding indicates that lambda repressor behaves in accord with that suggestion. Thus, as already mentioned lambda repressor, working at its own promoter with wild-type polymerase, primarily stimulates the isomerization step⁹⁵, but with a polymerase bearing a single amino-acid change (in σ^{70}), the effect is primarily on the initial binding step⁹⁷. For other examples in which similar activator-polymerase interactions are believed to have context-dependent effects, see refs 98-100. Activator-polymerase interactions that affect some step beyond recruitment—such as that between NTRC and polymerase—would be expected to be qualitatively different from those that simply recruit. The fact that activation by NTRC requires ATP but that by lambda repressor and CAP does not, is consistent with that expectation²⁷⁻²⁹.

body of evidence shows that the typical activator works by binding to specific sites on DNA and contacting RNA polymerase.

Recruitment

We begin with a discussion of two *Escherichia coli* activators—CAP (for catabolite activator protein), and the phage-lambda repressor working in its guise as an activator—that help polymerase bind stably to promoter DNA. This characterization (see Box 1), identifies these activators as working by recruitment, and distinguishes their action from that of another bacterial activator described below. A generalization that emerges from a large body of work is that any of a wide array of activator-polymerase interactions (Fig. 3), including a totally artificial one, can effect recruitment, and multiple interactions can work synergistically. The following sections summarize results of experiments supporting this generalization.

Activator-polymerase interactions

Lambda-repressor dimers bind cooperatively to adjacent DNA sites, one of which positions repressor very near the polymerase binding

site at the promoter (called P_{RM}) of the repressor gene². At this position, repressor contacts polymerase and thereby activates transcription of that gene. The following genetic experiments indicate that repressor activates by contacting the σ (in this case σ^{70}) subunit of polymerase. Repressor mutants specifically deficient in the activation function (called PC for positive control) have been described³⁻⁵; these mutants bear single amino-acid substitutions on their surface that most closely approaches RNA polymerase when both proteins are bound at the promoter. Certain mutations in σ^{70} specifically reduce the response to wild-type repressor⁶, whereas another restores response to a PC mutant repressor⁷.

CAP, in the presence of the cofactor cyclic AMP, binds to and activates several promoters, including those of the *Lac* and *Gal* genes. CAP contacts a subunit of polymerase different from that contacted by lambda repressor—namely α —and the precise interaction differs at different promoters^{8,9}. When bound very near polymerase (as at the *Gal* promoter), CAP makes two contacts with α using two different activating surfaces: one with the amino part of α and the other with the carboxyl domain of

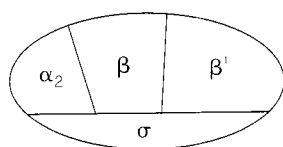


Figure 2 Each of the essential four subunits of *E. coli* RNA polymerase is present as a single copy, except for α of which there are two. The σ subunit directs the enzyme to specific promoters. Thus σ^{70} , the most commonly used form, recognizes characteristic sequences centred around positions 10 and 35 base pairs upstream of the transcription start site. The enzyme bearing σ^{54} recognizes different promoter sequences.

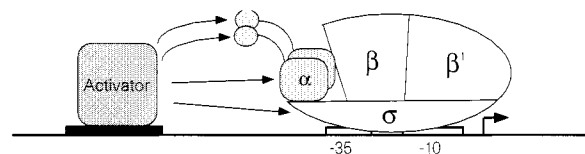


Figure 3 Activators can contact any of a variety of sites on RNA polymerase to recruit the enzyme and trigger gene activation. As discussed in the text, the precise contact or contacts made depend upon the identity of the activator and its position on DNA. The α -subunit is shown in greater detail than in Fig. 2. The α carboxy-terminal domain, a peptide of some 80 amino acids, is linked to the remainder of the molecule by a flexible linker.

that subunit¹⁰. When positioned upstream of polymerase (for example, 10 base pairs as at the Lac promoter), CAP contacts only the C-terminal domain of α . CAP need not work only through α : a mutant CAP has been described that replaces an α contact by a new contact with σ^{70} (refs 8, 11). As with lambda repressor, these contacts have been defined genetically: PC mutations have been isolated in both activating regions of CAP^{10–12}, and specific mutations in α abolish interaction with CAP^{13–15}. Chemical cross-linking provides additional evidence for these CAP–polymerase interactions^{10,16}. At least 12 activators in addition to CAP touch the α carboxyl domain¹⁷, and each of these activators contacts one of several different sites on this domain.

The fact that, as we have seen, a variety of different activator–polymerase interactions can activate transcription suggests that no special protein–protein interaction is required for recruitment. This idea is supported by two further observations. First, some of these activator–polymerase interactions can be replaced by a protein–DNA interaction: the promoters of ribosomal RNA genes work at a high level in the absence of activators, an effect attributable to a DNA ‘up’ element that is specifically recognized by the α carboxyl domain¹⁸. (For an example in which a polymerase–activator interaction compensates for a loss of a protein–DNA interaction, see ref. 20.) Second, as we will now describe, transcription can be activated by a protein–protein interaction not normally involved in gene activation.

Artificial recruitment

The following experiment²¹ shows that at promoters activated by both CAP and lambda repressor, the various activator–polymerase interactions described above can be replaced by a heterologous protein–protein interaction. This is an example of an ‘activator bypass’ experiment, in which activation is effected in the absence of a natural activating region (Fig. 4); similar experiments will play an important role in our analysis of activation of eukaryotic genes as well. In this instance, the carboxyl domain of α (present in two copies in the polymerase) was replaced by the carboxyl domain of lambda repressor, a domain that can form dimers and tetramers (Fig. 4a). When a lambda repressor dimer was bound upstream (at a position where it cannot activate using the natural activating region located on its amino domain), interaction between the carboxyl domains of that repressor and the carboxyl domains displayed by the modified polymerase (that is, tetramerization) triggers gene activation (Fig. 4b). Moreover, among a series of lambda mutants, the degree of tetramerization, measured in a separate experiment, roughly predicted the degree of activation.

Synergy

Other experiments in bacteria reveal an additional important principle of activation: multiple (in these cases two) activator–polymerase contacts have synergistic effects. That is, the level of transcription elicited by two contacts is significantly greater than the

sum of the levels elicited by each single contact. One example uses the experimental system just described in which polymerase bears the modified α subunit. When lambda repressor is positioned very near polymerase (as it ordinarily is at P_{RM}), it can make two contacts: the natural contact between its amino domain and σ^{70} , and the artificial contact involving the lambda repressor tetramerization interaction. Together these two interactions elicit a much greater level of activation than either single interaction²¹. In this case (as in the previously discussed case of CAP working at the Gal promoter), the two contacts are made by a single DNA-bound activator. A synergistic effect can also be achieved by two DNA-bound activators each making a single contact. In one example, one molecule of CAP and another of lambda repressor work synergistically on an artificial construct²². In another, two CAP molecules (both in an artificial and a natural setting) work synergistically^{23,24}.

A contrasting example

A family of bacterial activators has been described whose members, unlike CAP and lambda repressor, work at promoters that bear stably prebound polymerase. In *E. coli*, for example, polymerase (bearing σ^{54}) binds stably but inertly to the promoter of the *glnAp2* gene^{25,26}. As described in Box 1, the activator NTRC converts the complex to a form capable of initiating transcription. This effect, which, unlike activation by CAP and lambda repressor requires hydrolysis of ATP^{27–29}, evidently involves a more elaborate activator–polymerase interaction than those we have described so far. For example, NTRC might, by interacting with one or more special sites on polymerase, trigger a conformational change that stimulates the initiation of transcription. Another *E. coli* activator, the phage protein N4SSB, is also suspected to interact with polymerase that has already bound stably to DNA. In this case, the polymerase bears σ^{70} , and the activator target is the β' subunit³⁰. One interesting consequence of the fact that these activators work on prebound polymerases is that, when present at sufficiently high levels, they can activate without themselves binding to DNA^{30–32}, a matter we return to below.

Yeast

In yeast the transcriptional apparatus is more complicated than in bacteria, and in addition, genes are evidently less accessible because they are wrapped in histones. Nevertheless, we shall argue that, as in bacteria, any of a variety of activator–target interactions, including artificial ones, will activate transcription by recruiting the machinery to DNA. According to this picture, histones present a barrier against which polymerase recruitment is effected.

The transcriptional machinery

Most genes encoding messenger RNA are transcribed by one of three RNA polymerases, RNA polymerase II (Pol II), an enzyme comprising twelve subunits. Unlike the bacterial polymerase, Pol II requires many additional proteins—at least 30—to recognize a promoter

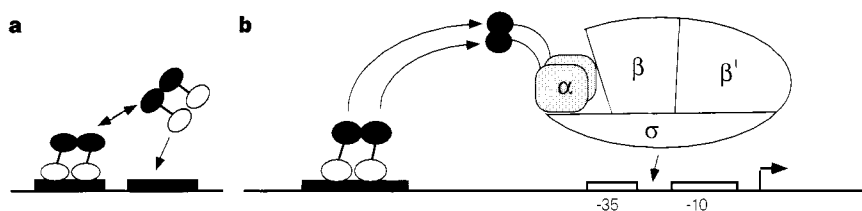


Figure 4 A heterologous protein-protein interaction can trigger gene activation in bacteria. **a**, Two lambda-repressor dimers bind DNA cooperatively by tetramerization of their carboxy termini, here represented as black circles. **b**, The tetramerization reaction can be used to recruit a hybrid polymerase in which the α carboxy-terminal domains are replaced by the carboxy-terminal domains of lambda repressor. In this experiment lambda repressor is bound some 20 base

and initiate transcription. Our knowledge of promoter structure would seem to be rudimentary: typical promoters are characterized as having an (A + T)-rich sequence (the TATA element) positioned some 50–70 base pairs upstream of the transcriptional start³³, and there is some evidence for initiator sequences around the start sites. Binding sites for yeast activators are typically found 100–250 base pairs upstream of the genes they regulate (for reviews, see refs 33 and 34).

The proteins that direct transcription, including PolII, are believed to be organized into a small number of complexes that work together^{35,36} (Fig. 5). This picture of the transcriptional machinery is a relatively new one that has arisen from the results of genetic and biochemical experiments. Early experiments, performed with mammalian cell extracts, revealed a series of general transcription factors (GTFs) which were required, in addition to polymerase itself, for promoter recognition and transcription initiation *in vitro*. At least some of these original factors contain more than one component: for example, transcription factor TFIID contains TBP, the protein that binds to the TATA sequence, and in addition a group of proteins called TAFs (for TBP-associated factors). The separated GTFs were found to assemble in a unique order at the promoter: TFIID first, followed by TFIIB, and then, in a defined order, the remaining factors attach (for a review, see ref. 37). Similar results were found by analysis of yeast cell extracts, and many of the components are interchangeable between the yeast and mammalian systems.

Genetic analysis in yeast revealed, however, that a group of proteins in addition to the GTFs are required for transcription *in vivo*. Thus, starting with a yeast strain bearing a damaged polymerase, Young and colleagues isolated mutations in nine different genes called *SRBs* (for suppressor of RNA polymerase B, the B referring to an alternative name for polymerase II) that restored normal growth to those cells³⁵. It was subsequently revealed that the SRB proteins can be isolated in a large complex that includes polymerase and certain other GTFs, including TFIIB^{38,39}. This holoenzyme (Fig. 5) bears all the factors known to be involved in transcription initiation except for TFIID, which we discuss in greater detail below, and TFIIE. The aggregation state of the holoenzyme *in vitro* depends upon the isolation conditions. For example, RNA polymerase can be separated from a 'mediator' subcomplex which contains certain SRB proteins, TFIIF and Gal11, a protein we will discuss later. There is therefore room for debate as to which is the physiologically relevant form of the holoenzyme, before DNA binding, *in vivo*. We have only limited knowledge of the roles played by many of the components of the transcriptional machinery. One case for which we have structural information involves TBP. This protein binds to and distorts DNA bearing the TATA sequence so that a platform is presented for subsequent binding of TFIIB⁴⁰. How most of the other components are arranged at the promoter is not known.

Despite the increased complexity of the transcriptional machin-

ery, the basic design of eukaryotic activators is similar to that of prokaryotic activators: each bears an activating region and a DNA-binding region. One salient difference, illustrated by typical yeast activators such as Gal4 (ref. 41), and GCN4 (ref. 42) is that the activating regions can be physically separated from the DNA-binding regions; each of these activating regions functions when tethered to DNA by fusion to a heterologous DNA-binding domain. Certain natural activators (such as the herpes virus protein VP16) do not themselves bear DNA-binding domains, but rather attach to DNA by contacting other DNA-binding proteins.

Activation

A series of recent activator bypass experiments demonstrate gene activation by recruitment in yeast. The experiments are similar in concept to that already discussed, in which an artificial protein-protein interaction between a DNA-tethered protein and RNA polymerase triggers gene activation in bacteria (Fig. 4). We begin this section by describing examples of such experiments in yeast, and then, in the light of these, consider the action of natural activators.

Artificial recruitment. The first activator bypass experiment we consider involves a protein-protein interaction, created by spontaneous mutation, that enables an otherwise inert DNA-tethered peptide to activate transcription (Fig. 6a). The experiment shows explicitly that interaction between a DNA-tethered protein and the transcriptional machinery can trigger gene activation. The mutation changes a single amino acid in Gal11, the component of the holoenzyme mentioned above. (Gal11's role in transcription is not limited to the *GAL* genes, and its name is a historical artefact; deletion of *GAL11* decreases transcription generally some five- to tenfold.) In an ordinary yeast strain, the dimerization region of Gal4 (residues 58–97) has no detectable transcriptional activating function when tethered to DNA. In contrast, in cells bearing the mutant *GAL11* (called *GAL11P* for potentiator), the Gal4 fragment works as an activating region (provided that it is tethered to DNA by fusion to a DNA-binding domain) nearly as powerfully as does that of Gal4 itself. *In vitro* Gal11P (but not Gal11) interacts with Gal4(58–97), and among a series of *GAL11P* alleles, the strength of this interaction is proportional to the level of gene activation observed *in vivo*. Peptide fragments bearing the interacting regions of these proteins can be swapped without loss of function. Overexpression of a fragment comprising either interacting component, unattached to a DNA-binding domain, inhibits activation, evidently because the fragment titrates its interacting partner. These and other experiments argue strongly against the possibility that the Gal11P/Gal4(58–97) interaction conveys some special conformational change that triggers activation; rather, they suggest that this arbitrary interaction activates gene expression to the extent of the binding energy involved, a result that is to be expected if the interaction serves to recruit the holoenzyme to DNA^{43,44}.

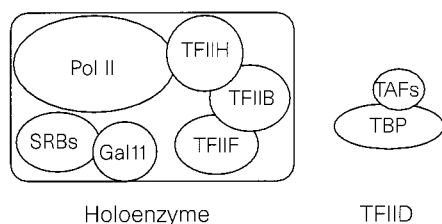


Figure 5 The yeast machinery that initiates transcription is organized in complexes. Some of the depicted components of the holoenzyme themselves comprise multiple proteins. For example, Pol II has 12 subunits and there are nine different SRBs; only some of the known holoenzyme components are shown. Not all of the Pol II is found in the holoenzyme, nor is all of the TBP complexed with TAFs, of which there are about 12. An essential factor that is not shown, TFIIE, is not strongly associated with either complex.

The Gal4(58–97)/Gal11P interaction can also drive gene activation in bacteria in an experiment similar to that of Fig. 4: in a strain bearing a polymerase fused through its α carboxyl domain to a fragment of Gal11, Gal4(58–97), tethered to DNA, activates transcription, provided the Gal11 fragment bears the P mutation²¹. Thus an identical protein–protein interaction will trigger gene activation in bacteria and in yeast.

The idea that the Gal4(58–97)/Gal11P interaction activates by recruiting the holoenzyme leads to the following prediction: fusion of a DNA-binding domain to a holoenzyme component would allow transcription at genes bearing the appropriate DNA binding site in the absence of any activator. This prediction is realized: yeast cells containing, for example, Gal11 fused to the DNA-binding domain of the bacterial protein LexA transcribe at a high level genes bearing LexA binding sites (Fig. 6b)⁴³. Various SRB proteins, including SRB2, SRB7, and SRB11, have all been shown to have a similar effect when fused to a DNA-binding domain (ref. 44 and L. Gaudreau, J. Nevado and M.P., unpublished results). In some cases, the level of gene activation is very high: LexA–Gal11, for example, works when the LexA binding sites are positioned quite far upstream (1,000 base pairs away for example) of the gene; when assayed at these large distances, LexA–Gal11 activates even more strongly than Gal4, itself a very efficient activator in yeast⁴³.

It is imagined that, in these experiments, each fusion protein is incorporated into the holoenzyme, and that the attached DNA-binding domain tethers the complex to DNA as depicted in Fig. 6b. Any required components that are not part of the holoenzyme (for example, TBP and TFIIE) would then bind cooperatively with it and trigger transcription. A similar picture would explain activation by TAF⁴⁵ and TBP^{46–48} fusions; in these cases the holoenzyme would bind cooperatively with TBP that is tethered directly or by association with a tethered TAF.

There is further evidence supporting the idea that the holoenzyme fusion proteins insert into the holoenzyme, as shown in Fig. 6b. A plausible interpretation of the genetic experiments of Koeleske *et al.*⁴⁹ is that certain SRB mutants more readily associate with a mutant polymerase (that used in the original selection for the SRB mutants) than do the wild-type forms of these SRB proteins. One such mutant is SRB2-1 (ref. 49) and, as expected on the basis of the scheme just described, LexA–SRB2-1 activates transcription in a strain bearing the damaged polymerase more efficiently than does a LexA fusion bearing the wild-type form of SRB2 (ref. 44). Also, holoenzyme has been isolated from a strain bearing, in place of wild-type Gal11, a Gal11–PHO4 fusion protein. In this fusion protein PHO4's activating region is replaced by Gal11. Template DNA bearing PHO4 binding sites is transcribed more efficiently than is template lacking those sites in a reaction containing, in

addition to the purified holoenzyme, purified TBP and TFIIE. (L. Gaudreau and M.P., manuscript in preparation).

Natural activation. We have seen that in yeast, as in bacteria, recruitment of the transcriptional machinery can trigger gene activation. We now consider whether natural yeast activators work in this way. The evidence suggests that activating regions can contact a variety of targets, and the recruitment model suggests that any of those interactions could trigger gene activation.

An early observation suggesting that the typical yeast activator does not work on a stably prebound transcriptional machinery arose from experiments in which activators that were damaged in their DNA-binding functions were expressed in cells at unusually high levels. To understand the implication of such experiments, consider the expectations based on the two models for gene activation we have discussed. First, consider the case in which the activator works on a polymerase stably bound to DNA. Ordinarily, activation requires that the activator binds DNA so as to raise its concentration to an appropriate level in the vicinity of the promoter. We would imagine that high activator concentrations would obviate that requirement, and the activator could work on the bound polymerase without itself binding to DNA. In fact, that is the effect observed with mutant forms of NTRC and with N4SSB, the bacterial activators that work on prebound polymerases^{29–31}. Second, consider activators that work by recruitment. We would anticipate that such activators could not activate were they unable to bind DNA, because in this case DNA tethering is required to bring the target to the DNA.

In fact, the typical result of overexpressing natural activating regions in yeast is inhibition, not activation. For example, overexpression of GAL4 inhibits activation by GCN4 of a gene bearing GCN4 sites⁵⁰. This 'squelching' phenomenon, widely observed with various activators in disparate organisms, is explained by assuming that the two activators have common target sites, and titration of those sites by one activator (at high concentration) prevents interaction with the second, DNA-bound, activator. We saw a demonstration of squelching (actually an example of 'self-squelching') in experiments with the Gal4(58–97)/Gal11P system⁴⁴. In that case, overexpression of Gal4(58–97), unattached to a DNA binding domain, inhibited activation by DNA-tethered Gal4(58–97).

Also consistent with the idea that yeast activators do not work on a prebound machinery are experiments showing that at several promoters activation is accompanied by marked changes in sensitivity to nucleases and chemical probes, a simple interpretation of which is that histones have been replaced by the transcriptional machinery^{51,52}.

Activating regions and their targets. Much of what we know about natural activating regions, otherwise quite puzzling, is at least partly explained by the assumption that their role is merely to stick to the transcriptional machinery and thereby recruit it to DNA. For example, no elaborate structure is required for the activating function⁵³. Anecdotal evidence suggests that typical activating regions—such as those found on Gal4 (ref. 41), GCN4 (ref. 54) and PHO4 (ref. 55), which bear an excess of negatively charged residues—are largely unstructured in the absence of an interacting partner. For both GCN4 and Gal4, fragments of the activating region function when tethered to DNA. In the Gal4 case, for example, a recent study found that, among a series of fragments extending from 17 to 41 amino acids, activation was approximately proportional to length⁵⁶. There must of course be some specificity determinants in activating regions that distinguish them from random peptides, but we do not have a clear idea as to what they are. The apparent lack of a structural requirement, however, would seem to be most readily explained by invoking simple interactions⁵⁷ that recruit. Some activating regions perform disparate functions with different sequence requirements. For example, the 41-amino-acid activating fragment of Gal4 alluded to above also recognizes the inhibitor Gal80 (ref. 58); introduced proline mutations have, in several instances, little effect on the activation function but strongly

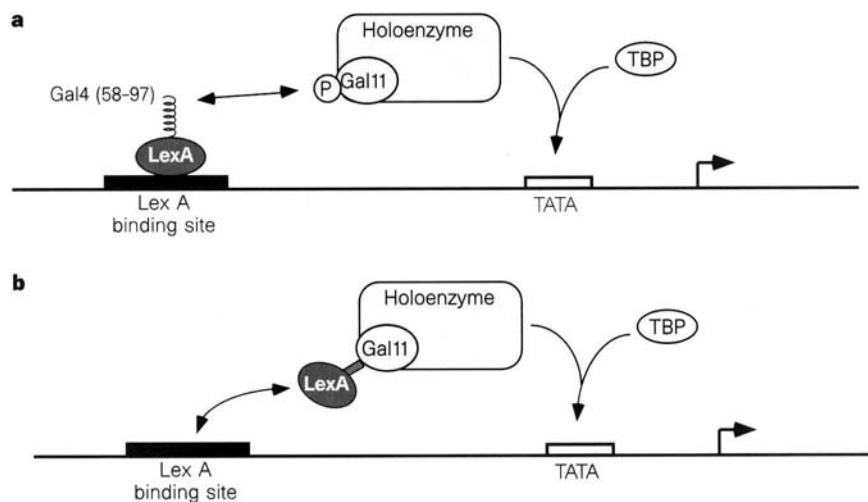


Figure 6 Gene activation in yeast can be effected in the absence of a natural activator by artificial recruitment of the holoenzyme. These are two examples of activator bypass experiments. **a**, The peptide Gal4(58-97), tethered to DNA by fusion to the DNA-binding domain of the bacterial repressor LexA, binds to the mutant holoenzyme component Gal11P and thereby recruits the holoenzyme to

the promoter. TBP binds cooperatively, presumably because it interacts both with DNA and with the holoenzyme. **b**, The LexA DNA-binding domain is fused directly to Gal11. The holoenzyme is recruited to DNA bearing a LexA-binding site near the beginning of a gene.

decrease Gal80 recognition, a result consistent with the idea that no very elaborate structure is required for the activation function (refs 41, 59, and A. Ansari and M.P., unpublished results).

An implication of this review is that, as in bacteria, there need be no specialized targets for activators; any of a large array of interactions would in principle suffice. In fact, many activator-putative target interactions have been detected *in vitro* in experiments with yeast and mammalian activators, although few of these have been quantified or extensively challenged with mutant forms of the interacting proteins⁵³. Showing relevance in any given case has been difficult, however, in part because the interactions may be redundant. Those proteins under serious consideration as potential activating region targets include the following.

TBP and TFIIB. An array of mammalian and yeast activators have been shown to interact with these proteins (for example, refs 56, 60, 61). In the quantitative study mentioned above, in which activation was proportional to the length of a fragment excised from Gal4, each individual fragment interacted indistinguishably with TBP and with TFIIB⁵⁶. Furthermore, the affinities of these interactions predicted the degree of gene activation. That relationship between affinity measured *in vitro* and activation measured *in vivo* was found to hold for a series of point mutants of activating fragments as well. A further suggestion that the measured affinities are physiologically relevant is provided by the observation that, for two unrelated interactions that give comparable levels of activation *in vivo* (that between the largest Gal4 fragment and TBP/TFIIB, and that between Gal4(58-97) and Gal11P), the measured affinities are approximately equal ($\sim 10^{-7}$ M) (refs 44, 56, 60). This energy of interaction is orders of magnitude greater than that measured for the bacterial activator CAP and *E. coli* RNA polymerase in the absence of DNA (R. H. Ebright, personal communication). That difference is what one would expect: the eukaryotic activators work when bound to sites positioned much further away from the transcriptional start than does this bacterial activator.

These and other results provide a strong, if not completely compelling, argument identifying TBP and TFIIB as activating region targets. At the same time they raise two questions, the resolution of which might require further analysis of the relevant structures. First, why is the affinity for either putative target protein approximately proportional to the length of the activating region

fragment? Might, for example, the activating region contain reiterations of some sequence pattern that is not evident upon inspection? Second, what common surface features of TBP and TFIIB dictate indistinguishable affinities for activator fragments in these experiments?

TAFs. It has been suggested that TAFs (TBP-associated factors) are required to transmit the effects of activators to the transcriptional machinery, and for this reason they were called coactivators. The simplest form of the recruitment model would seem not to require such specialized proteins and, as we have seen, such proteins are not required in bacteria. In addition to experiments already described, recent experiments suggest that in yeast as well there is no strong reason to assume that coactivators, in the sense defined here, are required. To understand this we must briefly review experiments that led to the coactivator hypothesis.

The proposed role of TAFs as coactivators was based on the findings that they were required for response to activators in mammalian and yeast cell extracts, and because they interacted with various activating regions. In those early experiments with crude extracts, removal of TFIID (that is, of TBP and TAFs) abolished transcription. Subsequent addition of purified TBP restored transcription, but that activity—called 'basal' transcription—was unaffected by activator⁶². In mammalian⁶² and subsequently in yeast experiments⁶³, response to activator was reestablished by adding a fraction containing TAFs to the reaction. These results were interpreted as showing that TAFs conveyed the activation signal from the activator to the basal transcriptional machinery. But subsequent analysis with mammalian extracts showed that a number of additional components—including PC4, an HMG protein, and other uncharacterized proteins—are also required for activation and are therefore by that criterion, co-activators⁶⁴.

Elimination of TAFs from yeast, although ultimately lethal, does not have a significant effect on activation of a wide array of genes^{65,66}. Therefore although the early *in vitro* results remain unexplained, if TAFs are targets of transcriptional activating regions in yeast, they are dispensable at most genes. Recent experiments with purified holoenzyme show that with at least one test promoter there is activation in the absence of TAFs³⁸.

In the TAF elimination experiments, although most genes were

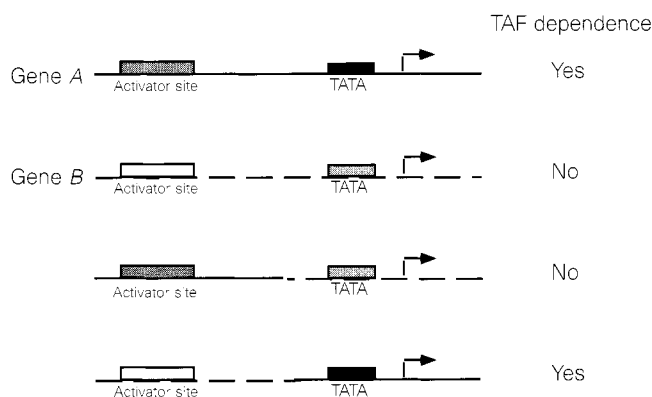


Figure 7 Certain genes cease transcription upon elimination of TAFs. This experiment shows that TAF dependence is determined by sequences around the TATA region and not by the nature of the activator. Each construct was assayed in dying yeast cells from which TAFs were eliminated as described⁶⁵. This experiment has been done with several pairs of genes. An example of gene *A* is *RPS5* and of gene *B* *ADH1* (W. Shen, S. Walker and M. Green, manuscript submitted).

unaffected, certain genes were found to cease transcription. Two scenarios might be imagined for the role of TAFs at those genes: they could be coactivators whose sole role is to transmit the activation signal, or they could be part of the transcriptional apparatus required at these promoters. That the latter is correct is suggested by the following experiment involving two genes, here called *A* and *B* (Fig. 7). The genes have upstream activating sequences that bind specific but different activators; transcription of *A* is TAF-dependent, but that of *B* is not. Hybrid genes were constructed in which the activator-binding elements were swapped. In the swapped configuration, the gene bearing the activating sequence from *A* was transcribed in a TAF-independent manner, whereas transcription of the gene bearing the corresponding sequence from *B* required TAFs. Thus, the TAF requirement of certain genes is specified by the sequence flanking the TATA element, and not by the nature of the activator. The simplest explanation for these results is that the TAF requirement arises not because certain activators must work through TAFs, but rather because at certain genes TAFs constitute an essential part of the transcription apparatus. This surmise is consistent with the finding that TAFs are DNA-binding proteins which could contribute binding energy required at certain promoters⁶⁷. This picture does not of course exclude the possibility that TAFs, like other components of the machinery, are activating-region targets. Evidence that TAFs might play such a role in higher eukaryotes has been presented⁶⁸.

In retrospect, an impetus for entertaining the coactivator hypothesis was the observation that, as we have seen, the machinery that directs activated transcription *in vitro* is apparently different from that which directs basal transcription. We suggest that this distinction is not physiologically relevant, and that basal transcription, should it occur in yeast, would best be described as we envisage basal transcription in bacteria. In bacteria, the term 'basal' transcription refers to the low level of transcription observed at certain promoters in the absence of activator. For example, the lambda-repressor gene, which is stimulated by lambda repressor, is transcribed at a low level *in vivo* in the absence of repressor. That basal transcription is understood to result from the spontaneous and infrequent stable binding of the intact polymerase to the promoter. A corresponding example of basal transcription in yeast may be that observed, in the absence of activator, when histones are removed from the cell (see below). We imagine that, as in bacteria, this basal transcription arises from the spontaneous stable binding of the intact transcriptional machinery to the promoter.

Others. There is no reason to exclude other components of the transcriptional machinery from consideration as activating region targets. For example, both the 'mediator' subcomplex of the holoenzyme and TFIIF interact with certain mammalian activators^{69,70} and perhaps these are also bona fide targets in mammalian and yeast cells.

Multiple interactions and synergy

We have seen that in bacteria multiple activator-polymerase interactions can work synergistically, and we would expect this to be true in yeast as well. The idea that yeast activating regions can touch multiple targets would provide an obvious way to achieve this effect, but there may be an additional consideration. In bacteria the entire machinery essential for the initiation of transcription is present in a single complex (the RNA polymerase) and so each contact simply adds energy to the binding reaction. However, as already pointed out, the yeast holoenzyme as currently isolated lacks at least one or two essential components, including TBP, and it is possible that, as suggested by Struhl³⁴, the different subcomplexes could be recruited independently. The following experiments are consistent with that view.

In the first experiment, Strubin *et al.*¹⁰¹ found that, in an artificial tethering experiment, a promoter bearing a weakened TATA sequence requires tethering of both the holoenzyme and, independently, TBP. In this case, recruitment of both components has a synergistic effect compared with the effect of recruitment of either component alone. In the second experiment, as noted earlier, at a wild-type promoter LexA-Gal11 is a stronger activator than Gal4. The reverse is the case, however, at a promoter bearing a damaged TATA sequence (A. Barberis and M.P., unpublished results). The results are plausibly explained by the observation that whereas LexA-Gal11 directly recruits only the holoenzyme, intact Gal4 molecules can touch both TBP and TFIIB.

It is unlikely that the activator-target interactions described here account for the full synergistic effects of activators. Experiments by Tanaka⁷¹, for example, suggest an important additional contribution from cooperative DNA binding. That effect evidently does not depend on direct interaction between the binding proteins, nor upon an indirect interaction mediated by interaction of activating regions with the transcriptional machinery (at least, not entirely). Rather, as he has proposed, some significant aspect of cooperativity may involve the concerted effect of two or more activators competing with histones for access to DNA (see also ref. 72).

Nucleosomes

The various examples of 'activator bypass' experiments described here suggest that holoenzyme recruitment is sufficient to overcome whatever obstacles to transcription are presented by nucleosomes and any other general inhibitors. An explicit test of this idea has been described using the *PHO5* gene in yeast, which is activated by PHO4. An array of nucleosomes is positioned at the promoter of this gene, and, as assayed by nuclease sensitivity, histones are removed or modified upon gene activation⁵⁹. When the activating region of PHO4 was replaced with a holoenzyme component (such as Gal11), the *PHO5* gene was efficiently transcribed and chromatin remodelled as usual⁷³. The experiment demonstrates that penetration of the nucleosomal barrier requires no special effect of an activator other than recruitment of the holoenzyme. Nucleosome remodelling is evidently a direct consequence of recruiting the transcriptional machinery: remodelling is observed even at a promoter deleted of its TATA sequence and thus unable to support transcription.

The simplest interpretation of these experiments, and of a related experiment performed with a yeast Pol III gene by Sentenac and colleagues⁷⁴, is that the holoenzyme competes for DNA access with histones. Histones would thus constitute an important part of the regulatory apparatus by presenting a barrier against which holoenzyme recruitment must be effected. Consistent with that idea is the

finding that histone depletion allows, to varying extents, spontaneous transcription of many genes⁷⁵. Weakening the histone barrier—by the binding of additional proteins and/or by histone modification—could therefore increase gene expression. Strengthening the barrier could decrease or even prevent transcription altogether. Widom and colleagues^{76,77} have measured the dissociation of histones from DNA *in vitro*, and on the basis of these results have proposed a model in which histones and regulatory proteins compete for common DNA sites⁷².

At least three protein complexes have been proposed to help in histone removal in yeast: SNF/SWI (ref. 78), GCN5/ADA (ref. 79) and RSC (ref. 80). One of these complexes (SNF/SWI) is reported (but see ref. 80) to be associated with the holoenzyme⁸¹, a finding that would account for the abilities of certain LexA–SNF fusion proteins to activate transcription⁸². One of these proteins, GCN5, is a histone acetylase, and active genes in higher eukaryotes have been characterized as being associated with hyperacetylated histones⁷⁹ (but see ref. 83). Mutation of components of the GCN5/ADA and SNF/SWI complexes has been reported to decrease expression of certain genes^{78,79}. In the experiments with the *PHO5* promoter described above, however, no significant effect on transcription was observed by mutation of either the SNF/SWI or GCN5/ADA complex. Full expression of the *PHO5* and *GAL1* promoters, activated by their natural activators, is similarly unaffected by mutation of either component. At the *GAL1* promoter, however, weakening activation (by, for example, using a promoter bearing only two Gal4 sites in place of the usual four sites) renders transcription elicited by either natural or holoenzyme–fusion activators sensitive to mutation of SNF/SWI or ADA/GCN5 (refs 73, 78, and L. Gaudreau and M.P., unpublished results). It is possible that as the total DNA–protein and protein–protein interactions are weakened, a greater contribution is required from these accessory factors. Whatever these complexes may do, the fact that their effects can be observed in activator bypass experiments shows that they can be called into play simply by recruiting the holoenzyme.

Higher eukaryotes

It would be surprising if the mechanism described here for gene activation in bacteria and yeast were absent from higher organisms. Mechanisms other than the one emphasized in our review are sure to be found in higher eukaryotes and perhaps in yeast as well (see refs 84–86 for example). Transcription of many genes, particularly in higher eukaryotes, is dependent upon multiple physiological signals. The following discussion shows how the recruitment mechanism we have described can help effect this integration.

The typical yeast gene discussed here is instructed to be transcribed at a high level by only one or two extracellular signals. Thus a few copies of DNA-bound Gal4, freed of the inhibitor Gal80 by the presence of galactose (and the absence of glucose), maximally activate the *GAL* genes. In the simplest elaboration of the recruitment mechanism, multiple activator–machinery contacts would be required for efficient transcription, and these contacts would be provided by disparate DNA-binding activators, each of which is controlled by a different physiological signal. Binding sites for these various activators, suitably positioned, would allow these activators to work together (synergistically as we have described) to activate the gene. In a further refinement we might imagine that, as already suggested⁸⁷, certain activators in higher eukaryotes differ from Gal4 in that they would be specialized to recognize more limited sets of potential targets. According to that idea, only certain combinations of such activators would work together at certain promoters because, at those promoters, two or more subcomplexes of the transcriptional machinery would have to be recruited independently. Other eukaryotic activators might be designed, like Gal4, to work independently. It is clear, however, that strong activation in higher eukaryotes does not intrinsically require such a complexity. Thus, for example, VP16, the strong activator expressed by herpes

simplex virus, works very efficiently without help from other activators⁸⁷. □

Mark Ptashne is in the Department of Molecular and Cellular Biology, 7 Divinity Avenue, Harvard University, Cambridge, Massachusetts 02138, USA; Alexander Gann is in the Division of Biological Sciences, Lancaster University, Lancaster LA14YQ, UK. E-mail: mptashne@fas.harvard.edu; a.gann@lancaster.ac.uk

1. Ptashne, M. How eukaryotic transcriptional activators work. *Nature* **335**, 683–689 (1988).
2. Ptashne, M. *A Genetic Switch: Phage Lambda and Higher Organisms* (Cell and Blackwell Scientific, Cambridge, MA, 1992).
3. Guarente, L. *et al.* Mutant lambda phage repressor with a specific defect in its positive control function. *Proc. Natl. Acad. Sci. USA* **79**, 2236–2239 (1982).
4. Hochschild, A., Irwin, N. & Ptashne, M. Repressor structure and the mechanism of positive control. *Cell* **32**, 319–325 (1983).
5. Bushman, F. D., Shang, C. & Ptashne, M. One glutamic acid residue plays a key role in the activation function of lambda repressor. *Cell* **58**, 1163–1171 (1989).
6. Kuddell, N. & Hochschild, N. Amino acid substitutions in the -35 recognition motif of sigma 70 that result in defects in phage lambda repressor-stimulated transcription. *J. Bact.* **176**, 2991–2998 (1994).
7. Li, M., Moyle, H. & Susskind, M. M. Target of the transcriptional activating function of phage lambda cI protein. *Science* **263**, 75–77 (1994).
8. Busby, S. & Ebright, R. H. Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* **79**, 743–746 (1994).
9. Busby, S. & Ebright, R. H. Transcription activation at class II CAP-dependent promoters. *Mol. Microbiol.* (in the press).
10. Niu, W. *et al.* Transcription activation at class II CAP-dependent promoters: two interactions between RNA polymerase. *Cell* **87**, 1123–1134 (1996).
11. Bell, A. *et al.* Mutations that alter the ability of the *Escherichia coli* cyclic AMP receptor protein to activate transcription. *Nucleic Acids Res.* **18**, 7242–7250 (1990).
12. Zhou, Y., Zhang, X. & Ebright, R. H. Identification of the activating region of CAP: isolation and characterization of mutants of CAP specifically defective in transcription activation. *Proc. Natl. Acad. Sci. USA* **90**, 6081–6085 (1993).
13. Tang, H. *et al.* Location, structure and function of the target of a transcription activator protein. *Genes Dev.* **8**, 3058–3067 (1994).
14. Zhou, Q. *et al.* Holo-TFIID supports transcriptional stimulation by diverse activators and from a TATA-less promoter. *Genes Dev.* **6**, 1964–1974 (1992).
15. Igarashi, K. & Ishihama, A. Bipartite functional map for the *E. coli* RNA polymerase alpha subunit. *Cell* **65**, 1015–1022 (1991).
16. Chen, Y., Ebright, Y. W. & Ebright, R. H. Identification of the target of a transcription activator protein by protein–protein photocrosslinking. *Science* **265**, 90–92 (1994).
17. Ishihama, A. Protein–protein communication within the transcription apparatus. *J. Bact.* **175**, 2483–2489 (1993).
18. Ross, W. *et al.* A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262**, 1407–1413 (1993).
19. Blatter, E. *et al.* Domain organization of RNA polymerase alpha subunit: C-terminal 85 amino acids constitute a domain capable of dimerization and DNA binding. *Cell* **78**, 889–896 (1994).
20. Kumar, A. *et al.* Role of the sigma 70 subunit of *Escherichia coli* RNA polymerase in transcription activation. *J. Mol. Biol.* **235**, 405–413 (1994).
21. Dove, S. L., Joung, J. K. & Hochschild, A. Activation of prokaryotic transcription through arbitrary protein–protein contacts. *Nature* **386**, 627–630 (1997).
22. Joung, J. K., Koepf, D. M. & Hochschild, A. Synergistic activation of transcription by bacteriophage lambda cI protein and *E. coli* cAMP receptor protein. *Science* **265**, 1863–1866 (1994).
23. Busby, S. *et al.* Transcription activation by the *Escherichia coli* cyclic AMP receptor protein—receptors bound in tandem at promoters can interact synergistically. *J. Mol. Biol.* **241**, 341–352 (1994).
24. Scott, S., Busby, S. & Beacham, I. Transcriptional coactivation at the ANSB promoters— involvement of the activating regions of CPR and FNR when bound in tandem. *Mol. Microb. Biol.* **18**, 521–531 (1995).
25. Kustu, S. *et al.* Expression of sigma 54 (ntrA)-dependent genes is probably united by a common mechanism. *Microbiol. Rev.* **53**, 367–376 (1989).
26. Magasanik, B. The regulation of nitrogen utilization in enteric bacteria. *J. Cell. Biochem.* **51**, 34–40 (1993).
27. Popham, D. L. *et al.* Function of a bacterial activator protein that binds to transcriptional enhancers. *Science* **243**, 629–635 (1989).
28. Austin, S. & Dixon, R. The prokaryotic enhancer binding protein NTRC has an ATPase activity which is phosphorylation and DNA dependent. *EMBO J.* **11**, 2219–2228 (1992).
29. Wedel, A. & Kustu, S. The bacterial enhancer binding protein NTRC is a molecular machine: ATP hydrolysis is coupled to transcriptional activation. *Genes Dev.* **9**, 2042–2052 (1995).
30. Miller, A. *et al.* RNA polymerase beta' subunit: target for DNA-binding-independent transcriptional activation. *Science* **275**, 1655–1657 (1997).
31. Buck, M. & Cannon, W. Mutation in the RNA polymerase recognition sequence of the *Klebsiella pneumoniae* nifH promoter permitting transcriptional activation in the absence of NifA binding to upstream activating sequences. *Nucleic Acids Res.* **17**, 2597–2612 (1989).
32. North, A. & Kustu, S. Mutant forms of the enhancer-binding protein NtrC can activate transcription from solution. *J. Mol. Biol.* (in the press).
33. Struhl, K. Molecular mechanisms of transcriptional regulation in yeast. *Annu. Rev. Biochem.* **58**, 1051–1077 (1989).
34. Struhl, K. Chromatin structure and RNA polymerase II connection: implications for transcription. *Cell* **84**, 179–182 (1996).
35. Koleske, A. I. & Young, R. A. The RNA polymerase II holoenzyme and its implications for gene regulation. *Trends Biochem. Sci.* **20**, 113–116 (1995).
36. Carey, M. A holistic view of the complex. *Curr. Biol.* **5**, 1003–1005 (1995).
37. Zawel, L. & Reinberg, D. Common themes in assembly and function of eukaryotic transcription complexes. *Annu. Rev. Biochem.* **64**, 522–561 (1995).
38. Koleske, A. & Young, R. A. An RNA polymerase II holoenzyme responsive to activators. *Nature* **368**, 466–469 (1994).
39. Kim, T. K. *et al.* Effects of activation-defective TBP mutations on transcription initiation in yeast. *Nature* **369**, 252–255 (1994).
40. Nikolov, D. B. *et al.* Crystal structure of a TFIIB–TBP–TATA element ternary complex. *Nature* **377**, 119–128 (1995).
41. Ma, J. & Ptashne, M. Deletion analysis of GAL11.4 defines two transcriptional activating segments. *Cell* **48**, 847–853 (1987).

42. Hope, I. A., Mahadevan, S. & Struhl, K. Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. *Nature* **333**, 635–640 (1988).
43. Barberis, A. *et al.* Contact with a component of the polymerase II holoenzyme suffices for gene activation. *Cell* **81**, 359–368 (1995).
44. Farrell, S. *et al.* Gene activation by recruitment of the RNA polymerase II holoenzyme. *Genes Dev.* **10**, 2359–2367 (1996).
45. Apone, L. *et al.* Yeast TAF (II)90 is required for cell-cycle progression through G(2)/M but not for general transcription activation. *Genes Dev.* **10**, 2368–2380 (1996).
46. Chatterjee, S. & Struhl, K. Connecting a promoter-bound protein to TBP bypasses the need for a transcriptional activation domain. *Nature* **374**, 820–822 (1995).
47. Klages, N. & Strubin, M. Stimulation of RNA polymerase II transcription initiation by recruitment of TBP *in vivo*. *Nature* **374**, 822–823 (1995).
48. Xiao, H., Friesen, J. D. & Lis, J. T. Recruiting TATA-binding protein to a promoter: transcriptional activation without an upstream activator. *Mol. Cell. Biol.* **15**, 5757–5761 (1995).
49. Koleske, A. J. *et al.* A novel transcription factor reveals a functional link between the RNA polymerase II CTD and TFIID. *Cell* **69**, 883–894 (1992).
50. Gill, G. & Ptashne, M. Negative effect of the transcriptional activator GAL4. *Nature* **334**, 721–724 (1988).
51. Almer, A. *et al.* Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *EMBO J.* **5**, 2689–2696 (1986).
52. Giardina, C. & Lis, J. T. Dynamic protein–DNA architecture of a yeast heat shock promoter. *Mol. Cell. Biol.* **15**, 2737–2744 (1995).
53. Treizenberg, S. J. Structure and function of transcriptional activation domains. *Curr. Opin. Gen. Dev.* **5**, 190–196 (1995).
54. Hope, I. A. & Struhl, K. Functional dissection of a eukaryotic transcriptional activator protein, GCN4 of yeast. *Cell* **46**, 885–894 (1986).
55. Ogawa, N. & Oshima, Y. Functional domains of a positive regulatory protein, PHO4, for transcriptional control of the phosphatase regulon in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **10**, 2224–2236 (1990).
56. Wu, Y., Reece, R. J. & Ptashne, M. Quantitation of putative activator–target affinities predicts transcriptional activating potentials. *EMBO J.* **15**, 3951–3963 (1996).
57. Harrison, S. C. Peptide–surface association: the case of PDZ and PTB domains. *Cell* **86**, 341–343 (1996).
58. Ma, J. & Ptashne, M. The carboxy-terminal 30 amino acids of GAL4 are recognized by GAL80. *Cell* **50**, 137–142 (1987).
59. Leuther, K. K., Salmeron, J. M. & Johnson, S. A. Genetic evidence that an activation domain of GAL4 does not require acidity and may form a β sheet. *Cell* **72**, 573–585 (1993).
60. Stringer, K. F., Ingles, C. J. & Greenblatt, J. Direct and selective binding of an acidic transcriptional activation domain to the TATA-box factor TFIID. *Nature* **345**, 783–786 (1990).
61. Nerlov, C. & Ziff, E. B. CCAAT/enhancer binding protein amino acid motifs with dual TBP and TFIIB binding ability co-operate to activate transcription in both yeast and mammalian cells. *EMBO J.* **14**, 4318–4328 (1995).
62. Pugh, B. F. & Tijan, R. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell* **61**, 1187–1197 (1990).
63. Reese, J. C. *et al.* Yeast TAFII in a multisubunit complex required for activated transcription. *Nature* **371**, 523–527 (1994).
64. Burley, S. K. & Roeder, R. G. Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* **65**, 769–799 (1996).
65. Walker, S. S. *et al.* Transcription activation in cells lacking TAFII. *Nature* **383**, 185–188 (1996).
66. Moqtaderi, Z. *et al.* TBP-associated factors are not generally required for transcriptional activation in yeast. *Nature* **383**, 188–191 (1996).
67. Verrijzer, C. *et al.* Binding of TAFs to core elements directs promoter selectivity by RNA polymerase II. *Cell* **81**, 1115–1125 (1995).
68. Sauer, F., Hansen, S. & Tijan, R. Multiple TAFII directing synergistic activation of transcription. *Science* **270**, 1783–1788 (1995).
69. Hengartner, C. J. *et al.* Association of an activator with an RNA polymerase II holoenzyme. *Genes Dev.* **9**, 897–910 (1995).
70. Xiao, H. *et al.* Binding of basal transcription factor TFIID to the acidic activation domains of VP16 and p53. *Mol. Cell. Biol.* **14**, 7013–7024 (1994).
71. Tanaka, M. Modulation of promoter occupancy by cooperative DNA-binding and activation-domain function is a major determinant of transcriptional regulation by activators *in vivo*. *Proc. Natl Acad. Sci. USA* **94**, 4311–4315 (1996).
72. Felsenfeld, G. *et al.* Chromatin structure and gene expression. *Proc. Natl Acad. Sci. USA* **93**, 9384–9388 (1996).
73. Gaudreau, L. *et al.* RNA polymerase II holoenzyme recruitment is sufficient to remodel chromatin at the yeast PHO5 promoter. *Cell* (in the press).
74. Marsolier, M. *et al.* Reciprocal interferences between nucleosomal organization and transcriptional activity of the yeast SNR6 gene. *Genes Dev.* **9**, 410–422 (1995).
75. Han, M. & Grunstein, M. Nucleosome loss activates yeast downstream promoters *in vivo*. *Cell* **55**, 1137–1145 (1988).
76. Polach, K. J. & Widom, J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation. *J. Mol. Biol.* **254**, 130–149 (1995).
77. Polach, K. J. & Widom, J. A model for the cooperative binding of eukaryotic regulatory proteins to nucleosomal target sites. *J. Mol. Biol.* **258**, 800–812 (1996).
78. Peterson, C. L. & Tamkun, J. W. The Swi–Snf complex: a chromatin remodeling machine? *Trends Biochem. Sci.* **20**, 143–146 (1995).
79. Brownell, J. E. & Allis, C. D. Special HATs for special occasions: linking histone acetylation to chromatin assembly and gene activation. *Curr. Opin. Genet. Dev.* **6**, 176–184 (1996).
80. Cairns, B. *et al.* RSC, an essential, abundant chromatin-remodeling complex. *Cell* **87**, 1249–1260 (1996).
81. Wilson, C. J. *et al.* RNA polymerase II holoenzyme contains SWI/SNF regulators involved in chromatin remodeling. *Cell* **84**, 235–244 (1996).
82. Laurent, B. C., Treitel, M. A. & Carlson, M. Functional interdependence of the yeast SNF2, SNF5, and SNF6 proteins in transcriptional activation. *Proc. Natl Acad. Sci. USA* **88**, 2687–2691 (1991).
83. DeRobertis, F. *et al.* The histone deacetylase RPD3 counteracts genomic silencing in *Drosophila* and yeast. *Nature* **384**, 589–591 (1996).
84. Greenblatt, J., Nodwell, J. & Mason, S. Transcriptional antitermination. *Nature* **364**, 401–406 (1993).
85. Benjtle, D. Regulation of transcriptional elongation by RNA polymerase II. *Curr. Opin. Genet. Dev.* **5**, 210–216 (1995).
86. Rasmussen, E. B. & Lis, J. T. Short transcripts of the ternary complex provide insight into RNA polymerase II elongation pausing. *J. Mol. Biol.* **252**, 522–535 (1995).
87. Blau, J. *et al.* Three functional classes of transcriptional activation domains. *Mol. Cell Biol.* **16**, 244–255 (1996).
88. McClure, W. R. Mechanism and control of transcription initiation in prokaryotes. *Annu. Rev. Biochem.* **54**, 171–204 (1985).
89. Ninfa, A. J., Reitzer, L. J. & Magasanik, B. Initiation of transcription at the bacterial glnAP2 promoter by purified *E. coli* components is facilitated by enhancers. *Cell* **50**, 1039–1046 (1987).
90. Sasse-Dwight, S. & Gralla, J. D. Probing the *Escherichia coli* glnA1G upstream activation mechanism *in vivo*. *Proc. Natl Acad. Sci. USA* **85**, 8934–8938 (1988).
91. Straney, D., Straney, S. & Crothers, D. Synergy between *Escherichia coli* CAP protein and RNA polymerase in the lac promoter open complex. *J. Mol. Biol.* **206**, 41–57 (1989).
92. Zhang, L. & Gralla, J. D. Micrococcal nuclease as a probe for bound and distorted DNA in lac transcription and repression complexes. *Nucleic Acids Res.* **17**, 5017–5028 (1989).
93. Sasse-Dwight, S. & Gralla, J. D. KMnO₄ as a probe for lac promoter DNA melting and mechanism *in vivo*. *J. Biol. Chem.* **262**, 8074–8081 (1989).
94. Malan, T. P., Buc, H. & McClure, W. R. Mechanism of CRP–cAMP activation of the lac operon transcription initiation activation of P1 promoter. *J. Mol. Biol.* **157**, 493–525 (1982).
95. Hawley, D. K. & McClure, W. R. Mechanism of activation of transcription initiation from the lambda-PRM promoter. *J. Mol. Biol.* **157**, 493–525 (1982).
96. Meyer, B. J. & Ptashne, M. Gene regulation at the right operator (OR) of bacteriophage lambda III: lambda repressor directly activates gene transcription. *J. Mol. Biol.* **139**, 195–205 (1980).
97. Li, M., McClure, W. & Susskind, M. Changing the mechanism of transcriptional activation by phage lambda repressor. *Proc. Natl Acad. Sci. USA* (in the press).
98. Smith, T. & Sauer, R. Dual recognition of open-complex formation and promoter clearance by Arc explains a novel repressor to activator switch. *Proc. Natl Acad. Sci. USA* **93**, 8868–8872 (1996).
99. Monsalve, M. *et al.* Activation and repression of transcription at two different phage E29 promoters are mediated by interaction of the same residues of regulatory protein p4 with RNA polymerase. *EMBO J.* **15**, 383–391 (1996).
100. Choy, H. *et al.* Repression and activation of transcription by Gal and Lac repressors: involvement of alpha subunit of RNA polymerase. *EMBO J.* **14**, 4523–4529 (1995).
101. Couto, G., Klages, N. & Strubin, M. Synergistic and promoter-selective activation of transcription by recruitment of TFIID and TFIIB. *Proc. Natl Acad. Sci. USA* (in the press).

Acknowledgements. We thank the participants of a lively meeting held at Cold Spring Harbor in December 1996 at which many of the ideas discussed here were developed. In particular, and additionally, we thank A. Hochschild, R. Ebright, S. Johnson, Dawit Haile, A. Barberis, S. Kustu, K. Yamamoto, G. Gill, R. Losick, L. Gaudreau, T. Maniatis, K. Lynch, A. Ansari and M. J. Wright.

Evidence for deep mantle circulation from global tomography

R. D. van der Hilst*, S. Widiyantoro† & E. R. Engdahl‡

* Massachusetts Institute of Technology, Department of Earth, Atmospheric, and Planetary Sciences, Rm 54-514, Cambridge Massachusetts 02139, USA

† Australian National University, Research School of Earth Sciences, Canberra ACT 0200, Australia

‡ US Geological Survey, DFC, MS 967, PO Box 25046, Denver, Colorado 80225, USA

Seismic tomography based on P-wave travel times and improved earthquake locations provides further evidence for mantle-wide convective flow. The use of body waves makes it possible to resolve long, narrow structures in the lower mantle some of which can be followed to sites of present-day plate convergence at the Earth's surface. The transition from subduction-related linear structures in the mid-mantle to long-wavelength heterogeneity near the core-mantle boundary remains enigmatic, but at least some slab segments seem to sink to the bottom of the mantle.

Global tomography, a technique pioneered in the 1970s and early 1980s^{1,2} to interpret the observed seismic wave field in terms of seismic properties at depth, is revolutionizing our knowledge of the Earth's interior. However, decades of research and spirited debate have not led to a consensus on the scale of mantle convection and the nature of coupling between motions in the deep interior and tectonic processes at the surface. Resolution of this issue is important for better understanding of the thermal and chemical evolution of our planet^{3,4}. There is increasing agreement on large-scale (>4,000 km) structural features in global images⁵⁻⁷, which represent the integrated effects of mantle processes over long periods of time^{8,9}. The shorter-wavelength components in the models, however, have appeared sensitive to the type of data and model parametrization used in the inversions and have so far not shown convincing correlation. This lack of agreement prevents the effective use of global models to constrain reconstructions of past plate motion and numerical flow simulations at high Rayleigh numbers, and to discriminate unambiguously between end-member circulation models known as layered-mantle and whole mantle convection. For recent reviews see refs 6, 10–13.

Several lines of geophysical evidence support flow across the boundary between the upper and lower mantle, defined here by the seismic discontinuity at approximately 660 km depth. The magnitude of lateral variations in depth to this interface¹⁴ is consistent with an endothermic phase change in the (Mg, Fe)₂SiO₄ system¹⁵ that is by itself too weak to stratify flow¹⁶. Global tomography models lack the increased heterogeneity near 660 km depth expected for a thermal boundary layer at that depth¹⁷, anomalous travel times of high-frequency waves are consistent with slab penetration into the lower mantle in some subduction systems^{11,18–25}, and long-wavelength gravity data can be explained by dynamic models of deep circulation in a high-viscosity lower mantle^{5,12,26}. On the other hand, regional seismological studies have demonstrated stagnation of some slabs in the upper-mantle transition zone^{20,21}. In addition, the long-term survival of distinct chemical reservoirs of primordial mantle as deduced from isotope and trace-element data has been used as evidence for layered convection (for reviews see refs 3 and 4). However, these data are not inconsistent with mantle-wide distribution of compositional heterogeneity^{12,27}, for which there seems to be increasing observational seismological evidence^{28,29}.

Here we present results of the inversion of carefully processed travel times, mainly from first-arriving P-waves, for aspherical variations in compressional wave speed in the Earth's mantle. The uneven sampling limits the resolution of aspherical structure in the upper mantle, but the new data reveal deep structure in unprece-

dented detail, which helps to assess the ultimate fate of the subducted slabs and their role in lower-mantle flow. We infer long, linear structures that locally connect to seismogenic slabs in the upper mantle and that divide the mid-mantle into large domains of shorter-wavelength heterogeneity. The transition from the linear subduction-related structures in the mid-mantle to long-wavelength heterogeneity near the core-mantle boundary (CMB) remains enigmatic, but our study reveals segments of slab that connect to structural heterogeneity in the CMB region. Along with results of regional studies that reveal structural complexity in the transition zone, these findings suggest that the Earth's present-day convective regime is predominated by some form of whole mantle overturn and their intermittent mantle stratification is a local³⁰ and transient^{31,32} phenomenon only.

Data and tomographic method

Shear-wave models are often based on carefully processed digital waveforms and can exploit the excellent spatial data coverage provided by a combination of direct and surface-reflected body²² and surface waves^{33–35} and by constraints of free oscillation data^{7,36}. In contrast, global P-wave models are typically based on relatively noisy short-period travel-time data reported to international data centres by thousands of station operators worldwide^{1,2,5,21,37,38}. An important novel aspect of our global imaging is the use of better travel-time data and earthquake locations than previously available.

In a major effort, Engdahl and co-workers³⁹ reprocessed the entire database published by the International Seismological Centre (ISC) and, for recent years, the US Geological Survey's National Earthquake Information Center. Using robust statistics, an improved global travel-time model⁴⁰, and the arrival times of direct P- and S-phases, depth phases (pP, sP and the ocean-surface-reflected pwP), and PKP core phases, they relocated in an iterative, nonlinear procedure all teleseismically well-constrained earthquakes occurring between 1 January 1964 and 31 December 1995 (Fig. 1a). Upon application of appropriate corrections and event selection criteria, this procedure minimizes focal depth errors and the mapping of source heterogeneity into mislocation, thereby creating a significantly improved database for tomographic imaging. We used an iterative, conjugate-gradient algorithm⁴¹ to invert all available P and pP travel-time residuals for effects of source mislocation due to three-dimensional structure and for P-wave speed in about 300,000 constant velocity blocks. We did not account for effects of aspherical structure on path geometry. This may underestimate the amplitude and overestimate the width of the anomalies in the upper mantle but does not effect the conclusions based on relatively large-scale lower-mantle structures.

The Earth's mantle is probably heterogeneous at all scales but the choice of a particular parametrization sets a minimum resolution length and excludes smaller structure from observation. Models based on spherical harmonics^{33–36} constrain long-wavelength anomalies, but do not resolve in detail trajectories of mantle flow such as descending slabs. The local basis functions used in our study (constant-velocity blocks with a dimension of $2^\circ \times 2^\circ \times 200$ km in the lower mantle) begin to provide the resolution required to investigate the continuity of structure between upper and lower mantle and the pattern of convective flow at larger depth. However, the uneven source/receiver distribution (Fig. 1a) and the effective restriction of inversions for P-wave speed to body waves necessarily renders large regions where the solution is not constrained, in particular in the upper mantle and transition zone (Fig. 1b). Body-wave sampling improves significantly in the mid-mantle (Fig. 1c–e) but degrades again in the lowermost mantle, especially in the Southern Hemisphere (Fig. 1f). Here we focus on the part of the solution pertinent to lower-mantle structure; detailed images of slabs in the upper mantle are better obtained by regional studies or global studies that allow for variable block size, which would ideally

be based on three-dimensional ray tracing to account for effects of ray bending.

Aspherical mantle structure

Our results for the upper mantle (not shown) are in general agreement with surface-wave studies^{6,7,33–35}; low wave speeds characterize marginal basins and tectonic continental regions, and zones of fast wave propagation outline stable continental cratons. The data reveal numerous fast anomalies in the upper-mantle transition zone beneath the circumpacific region (Fig. 1b), which agrees with inferences from Earth's free oscillations³⁶.

In the top half of the lower mantle, inversion of our improved data brings out prominent high-wave speed structures that continue intermittently over horizontal distances in excess of 10,000 km beneath the Americas and the southern margin of Eurasia (Fig. 1c, d). These outstanding structures correlate with anomalies in models based on spectral techniques (Fig. 2) but are narrower than inferred from these previous results, in particular at depths less than 1,500 km. With a width of 500–1,000 km in map view their shape is not controlled by the block size used. We infer that the

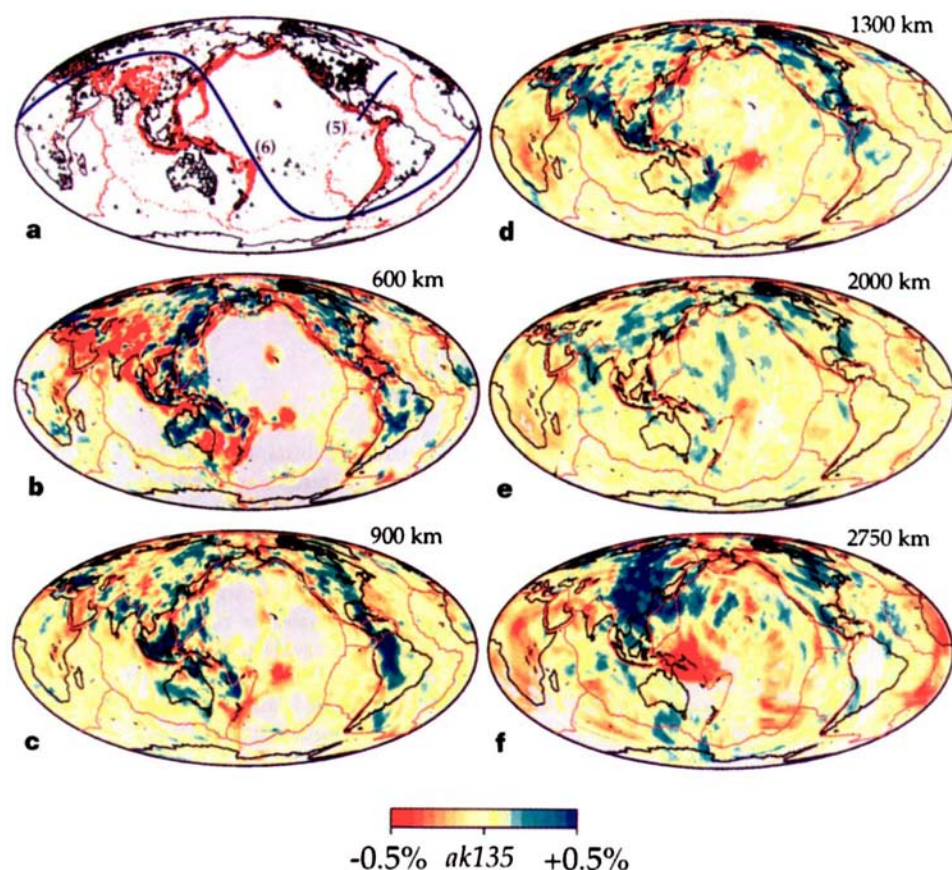


Figure 1 a, Pacific-centred Mollweide projection depicting the locations of the almost 80,000 earthquakes (red dots) and 3,500 stations (triangles) that produced the 7.3 million P and 0.3 million pP phases used in the inversions resulting in the P-wave-speed anomalies as displayed at 600 km (b), 900 km (c), 1300 km (d), 2,000 km (e) and 2,700 km (f) depth in the Earth's mantle. Dark blue lines in a depict the locations of the cross-sections of Figs 5 and 6. To suppress noise and reduce the spatial imbalance in data coverage we grouped data associated with event and station clusters into summary rays²⁵, which reduced the total number of data used in the inversions from 7.6 million to 500,000. Wave speed variations are relative to the radially stratified ak135 model⁴⁰; see colour scale. Blue (red) colours represent fast (slow) wave propagation; grey depicts mantle regions of poor sampling. In these map views, no *a posteriori* smoothing or interpolation is used other than regridding from $2^\circ \times 2^\circ$ to $1^\circ \times 1^\circ$ blocks. The amplitude is, however,

poorly constrained and therefore not used in the discussions. The perturbations may underestimate the actual values because of (1) the effect of regularization (damping) of the inversion, (2) trade-offs with earthquake location (on inversion, the hypocentre relocation parameters can in principle absorb substantial signal; in practice this is limited by the effective control on focal depth by the depth phases³⁹), and (3) the neglect of ray bending effects. The regularization applied is a combination of norm and gradient damping; in absence of strong data constraints the former produces a bias towards the reference values (that is, zero anomalies) and the latter minimizes the difference in amplitude between adjacent blocks, which results in a smooth model. The model shown reduces the variance of the summary ray data by almost 50%, in addition to the ~15% variance reduction pre-processing of the original ISC data^{20,39}. The thin red lines depict the location of plate boundaries.

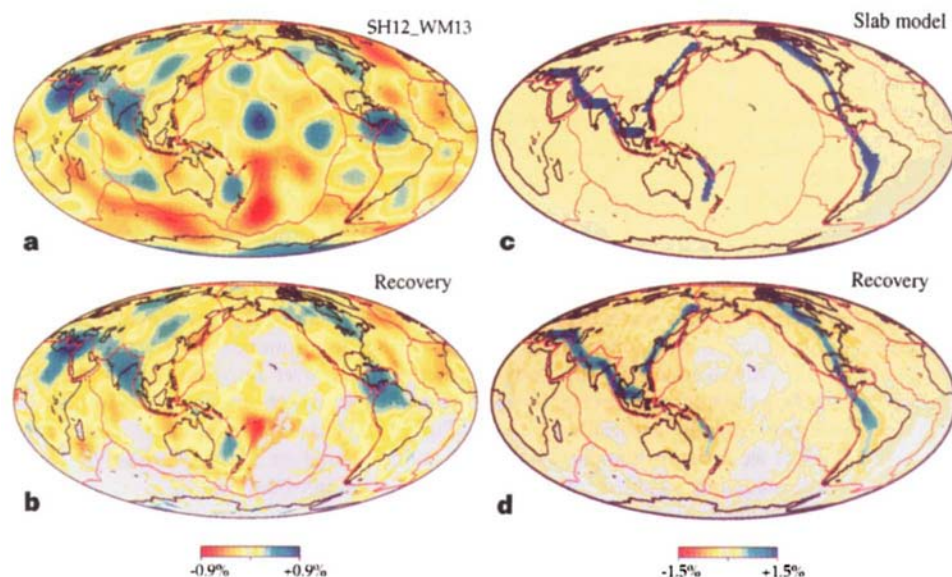


Figure 2 Resolution of long-wavelength (**a**, **b**) and short-wavelength (**c**, **d**) structure. We investigated the potential resolution of certain structural features in aspherical Earth models by inverting synthetic data calculated from assumed input models. We added gaussian noise to the synthetic data to simulate data errors and applied the same parameters (number of iterations, amount of regularization) as in the inversion of reported phase data. **a**, Lateral variation in shear-wave speed in the mid-mantle (at 1,300 km depth) according to *SH12_WM13* by Su and co-workers³³, and **b**, structure according to this model as it would have been imaged using the body-wave paths used in our study. Comparison of the input (**a**) and the recovery after inversion (**b**) demonstrates first, that our data coverage is sufficient, in particular in the Northern Hemisphere, to resolve large-scale structural features, and second, that there is significant agreement between the models if structures in *SH12_WM13* represent long-

wavelength images of narrower structures (Figs 1d and 3). **c**, Artificial slab structure at 1,300 km depth with a peak anomaly of 3% (off-scale). The result of the inversion of the (noisy) synthetic data (**d**) demonstrates that narrow structure in the lower mantle can be resolved by our data, in particular in the Northern Hemisphere, and that the width of the anomaly may not be significantly over-estimated. These tests also indicate that insufficient sampling by our body waves precludes the resolution of lower-mantle structures in large parts of the Southern Hemisphere. (Note that our colour coding intentionally de-emphasizes structure of near-zero amplitude. Inversion with random noise and the assessment of effects of, for instance, source mislocation and uneven data coverage suggest that structure with near-zero amplitude is not necessarily indicative of Earth's structure).

lower-mantle structure beneath the Americas is laterally coherent for depths up to approximately 1,700 km, but the slab signature vanishes at about 1,300 km depth beneath South America²². The deep anomaly connects to upper-mantle structures that are more fragmented as a result of Cenozoic tectonic activity in the eastern Pacific and Caribbean region. Beneath eastern Europe and Indonesia a fast anomaly can be traced from mid- to upper-mantle depths, but it is elsewhere only inferred between about 1,000 and at least 1,700 km in depth. The vertical dimension of both lineaments exceed the depth uncertainty in the images of about 300 km as determined by resolution tests with fabricated slab models (S.W. and R.D.v.d.H., manuscript in preparation). Outside these large-scale structures there is substantial scatter of smaller-scale heterogeneity, some of which may be real on the basis of correlation with independent results^{22,42}.

In the mid-mantle, slow anomalies appear isolated in map view. Their shape seems more equidimensional than the fast structures, but this is not well constrained owing to poor data coverage. A slow anomaly is visible from the Earth's surface to at least 2,000 km depth beneath the southwestern Pacific Ocean (Society islands) and from about 800 km depth to the CMB beneath southern and central Africa.

At about 1,700 km depth the character of heterogeneity begins to change: the linear structures (Fig. 1c, d) gradually disintegrate into smaller-scale anomalies that show less lateral continuity (Fig. 1e). In the lowermost mantle (2,300 km depth to the CMB), mantle structure is again dominated by long-wavelength features (Fig. 1f), but the general appearance of heterogeneity in this depth range is strikingly different from the linear structures at shallower depth

(compare, for instance, Fig. 1d, f). There is general agreement between our results and independent studies of the region just above the CMB (see Wyssession⁴³ and references therein), although there seem to be some differences between P- and S-wave speed that are not yet fully understood⁴². Our data do not satisfactorily resolve Southern Hemisphere structure but the global distribution of fast and slow anomalies is consistent with the heterogeneity pattern inferred from waveform data^{33–35}. The images also reveal small-scale features that we think are real. For example, the short-wavelength variations at mid-latitudes in the western Pacific are in excellent agreement with anomalous core-phase (PKP) times (K. C. Creager, personal communication).

Are narrow structures in the mid-mantle real?

The inference that narrow features are prominent to at least 1,700 km depth in the Earth's mantle is perhaps surprising as spherical-harmonic representations of the Earth's interior shear structure indicate that the amplitude of structure in the mid-mantle is significantly lower than that in the upper and lowermost mantle structure, at least out to degree 16 (refs 33–35).

The length of the linear features exceeds by far the distance over which body waves propagate horizontally, indicating that a large number of data with coherent structural signal must contribute to the mapping of these structures. The narrow width of the lower-mantle lineaments is not an artefact owing to preferential sampling or the use of local basis functions for model parametrization (Fig. 2). With the body-wave paths used in the actual data inversion we computed travel times using the long-wavelength model by Su and co-workers³³. We inverted these synthetic data and compared the

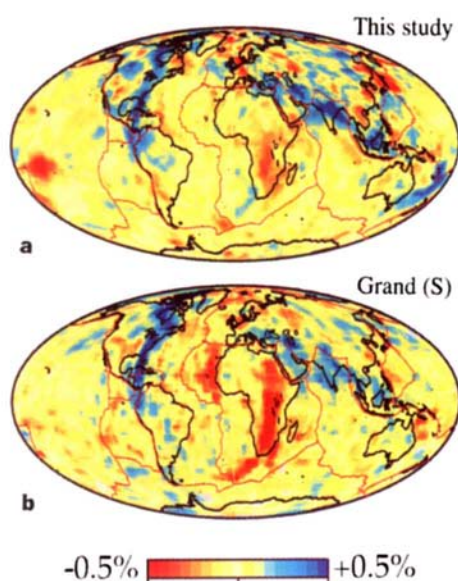


Figure 3 Africa-centred Mollweide projections of lateral variation in P- and S-wave speed at a depth of 1,350 km in different tomographic models of the Earth's mantle. The linear high-wave-speed anomalies revealed by inversion of our P-wave data (**a**) are in very good agreement with the results of S-wave inversions by Grand^{22,42} (**b**). Grand's model is based on a careful analysis of direct and multiply reflected shear waves, and is continually being upgraded with more data. The wave-speed perturbations are plotted on the same scale, that is between $\pm 0.5\%$ relative to the reference models used. For this comparison we choose to use a different projection than in Figs 1, 2 and 4 because mantle structure beneath the western Pacific is not yet well constrained in Grand's current models.

output to the input model. The excellent model recovery indicates that our inversion technique does not bias towards short-wavelength anomalies in regions of dense sampling and implies that long-wavelength lower-mantle structures will be mapped accurately by our method if the data contain signal pertinent to such structures. This interpretation, if correct, suggests that spectral methods map similar structural features but may significantly overestimate their true width. Our resolution degrades in areas of poor sampling and substantial parts of the long-wavelength models cannot be tested against our results. Such tests thus help isolate structures for which a model comparison is meaningful.

Resolution tests also demonstrate that narrow structures can be resolved in most parts of the lower mantle and that their width can probably be inferred fairly accurately from Figs 1 and 3 (Fig. 2c, d). The increasing agreement between independent models further suggests that the linear anomalies are real. Once identified, the linear structures can be recognized in the long-wavelength models^{33–35} (Fig. 2) and in images based on ISC data^{37,38} but these models typically show high-amplitude structure in other regions as well, which complicates interpretation. Particularly exciting is the excellent correlation between the narrow structures in our P-wave maps and similar features in the S-wave model by Grand^{22,42} (Fig. 3). There is good agreement also at depths other shown in Fig. 3 but differences exist in regions of poor data coverage in one or the other model, or both⁴². The spectacular correlation of structural detail in models deduced independently from different data and analysis procedures marks significant progress in global imaging.

Cold downwellings in the lower mantle

Figure 4a shows the lateral distribution (planform) of downwellings in the mid-mantle according to a synthesized three-dimensional

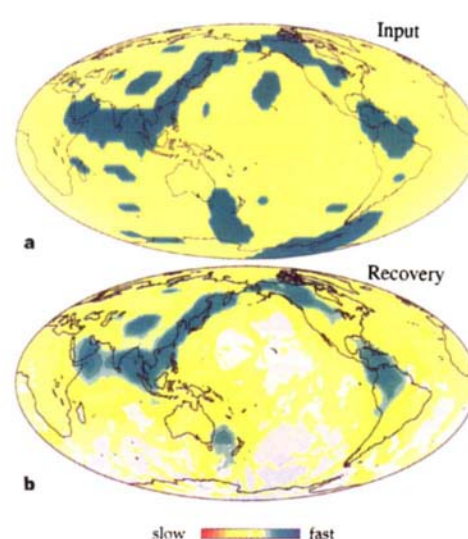


Figure 4 Illustration of the resolution of large-scale structures at depths between 1,800 and 2,000 km using a synthesized three-dimensional slab model⁹ that is based on plate convergence in the past 180 Myr (refs 8, 44). This model is produced by 'simply' dropping spherical particles, called slabs, into the mantle beneath a convergence margin and does not account for ambient mantle flow or dynamical effects of the transition zone. The 'slabs' sink vertically across the upper/lower mantle interface, but at a reduced rate due to an increase in viscosity. We inverted synthetic data calculated from the three-dimensional slab model, shown in **a** at a depth of 1,900 km, and compared the recovered model (**b**) to the input model. We tested the resolution of such large-scale structures at $\sim 1,900$ km because this represents the depth interval where inversion of the actual data indicates a patchwork of smaller scale structure. The comparison suggests that: (1) upon the damped inversion there is a substantial loss of amplitude; (2) our data coverage is not sufficient to resolve structure in the Southern Hemisphere and the central Pacific; and (3) the circumpacific and south Asia anomalies would be well resolved, but smearing can produce structures with amplitudes less than $\sim 20\%$ of the peak values. Tests such as those illustrated here and in Figs 2 and 5 have several advantages over more standard 'chess-board' tests. In particular, one can investigate more directly whether the data used can resolve a hypothetical pattern of structural heterogeneity. This technique can help, for instance, to explore the class of seismic data required to verify certain aspects in numerical models.

slab model⁹ based on plate convergence in the past 180 Myr (ref. 44) and a simple simulation of whole-mantle flow. Our data coverage is sufficient to map out the structure as predicted (Fig. 4), except for the region south of South America, which implies that many observed differences from the tomograms are meaningful.

The subduction model predicts significant slab structure in the mid-mantle beneath North America (subduction of the Farallon plate), southern Eurasia (Tethys ocean floor), the northwestern Pacific (Izanagi, Kula and Pacific plates) and Tonga (Pacific plate). The Farallon anomaly, the first lower-mantle structure associated with subduction⁴⁵ and discussed in more detail by Grand²², is prominent in our images and appears to be continuous from the upper mantle to the CMB beneath central America (Fig. 5). In the upper 300 km, the narrow slab does not leave a signature in $2^\circ \times 2^\circ$ blocks but shows up clearly when a smaller block size is used (H. Bijwaard *et al.*, personal communication). The subduction of former African lithosphere beneath Europe, often referred to as the Aegean slab, forms the western boundary of the Tethys anomaly in the mid-mantle beneath southern Eurasia (Fig. 6a). The Aegean slab seems to be continuous to approximately 1,500 km depth

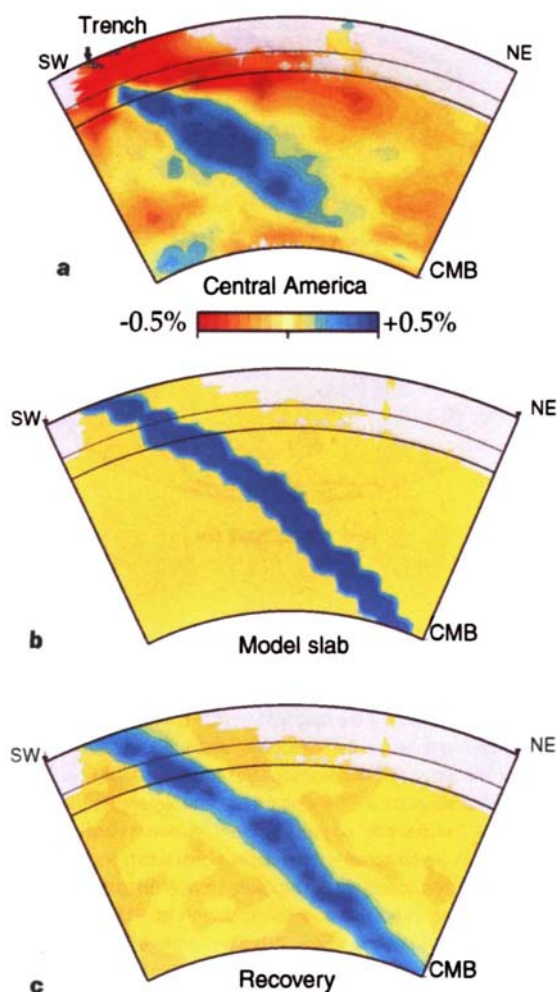


Figure 5 a, Vertical mantle section through our global P-wave model from the Earth's surface to the core-mantle boundary across the convergent margin in Central America (see Fig. 1a for cross-section location). This anomaly was one of the first lower-mantle structures interpreted as lithospheric slab in the lower mantle⁴⁶; see Lay¹³ for a recent review of this and other subduction-zone structures. The fast anomaly in the lower mantle has a peak value of almost 1% of the *ak135* reference model, but this value is not well constrained. As before, the poorly sampled mantle regions are masked. CMB, core-mantle boundary; NE and SW denote geographical orientation. The word 'trench' marks the location at the surface of the Middle America trench. Thin lines at constant depth depict the location of the 410 and 660 km discontinuities. The comparison of the input with a peak anomaly of 3% (**b**) and output (**c**) of test inversions with synthetic data (see Figs 2 and 4 for description) demonstrates that both the width and amplitude of the high-wave-speed anomaly beneath Central America are influenced by irregular sampling but that the slab is probably not much narrower than inferred from **a**. Results of studies at higher resolution (S.W. and R.D.v.d.H., manuscript in preparation) indicate continuity of the Farallon slab across the shallow mantle but that it is too narrow to be resolved in our current global inversion. Similar tests show that for Tonga the data can not discriminate between penetration to ~1,800 km or to a larger depth.

(Fig. 6b), which is in accord with a regional study²³. The eastern edge of the Tethys structure is marked by subduction of the Indo-Australian plate beneath the Sunda arc²⁵.

The pattern of mantle flow beneath the northwestern Pacific remains enigmatic. On the basis of plate reconstructions for this region^{8,9,44} one would expect a slab-related anomaly in the lower mantle that is as prominent as the Tethys and Farallon structures (Fig. 4). Even though our data coverage is sufficient to detect such a feature (Fig. 2c, d) we do not observe it in the depth range between approximately 700 and 1,400 km (Fig. 1c, d). This concurs with global models based on spherical harmonics: a circumpacific high is usually inferred, but in the top of the lower mantle beneath east Asia the fast anomaly is often either absent or significantly reduced in amplitude^{33–35} (Fig. 2; H. Bolten and G. Masters, personal communication). At approximately 1,500 km depth a large-scale anomaly re-emerges that continues to the CMB (Fig. 1). Locally, narrow fast anomalies seem to pierce through the 'gap' in the uppermost lower mantle and connect to (and, indeed, cause) the pronounced high-wave-speed anomaly in the CMB region beneath eastern Asia (Fig. 6b).

Deep circulation in the Earth's mantle

Figures 1, 5 and 6 suggest that fast anomalies are continuous across the '660-km' discontinuity to at least 1,700 km depth beneath many convergent margins and that the long narrow structures in the mid-mantle are related to subduction of former oceanic lithosphere, with some slab segments probably sinking all the way to the CMB. Lower-mantle flow is required to explain these observations—the distance

to the upper mantle is simply too large for them to be the result of conductive cooling alone—and must be causally related to upper-mantle flow unless the inferred spatial correlation of wavespeed anomalies is just coincidental. Flow across the seismic discontinuity offers the simplest explanation of the observed continuity of amplitude and geometry (for example, dip angle) of the anomalies. In particular, our results argue against worldwide mantle stratification by a flow-impeding interface at 660 km or stagnation of slab material in the lower mantle above 1,100 km (ref. 46). Penetrative convection—a hybrid flow model in which slab material flows back to the upper mantle after initial penetration into a chemically distinct lower mantle with slightly higher intrinsic density¹⁰—may explain our observations, in particular if the chemical boundary occurs at such a large depth that downwarping to the CMB is plausible or if large depressions of the boundary nucleate downwellings that control the planform of convection in the layer beneath them (mechanical coupling) and show up as fast anomalies. Either alternative is not easily distinguished from simple whole mantle flow on the basis of seismic imaging alone.

Local intermittence of flow into the lower mantle. Although inconsistent with worldwide stratification at 660 km depth the interplay of plate tectonic motions, radial increases in viscosity, and (perhaps) endothermic phase changes can locally distort mantle flow. Beneath the northwestern Pacific region, for example, some slabs deflect in the transition zone^{20,21} whereas others penetrate to larger depth^{18–20} (for example, Fig. 6b). Slab structure is here not well defined in the top of the lower mantle but re-emerges at larger depth (~1,500 km) and then continues to the

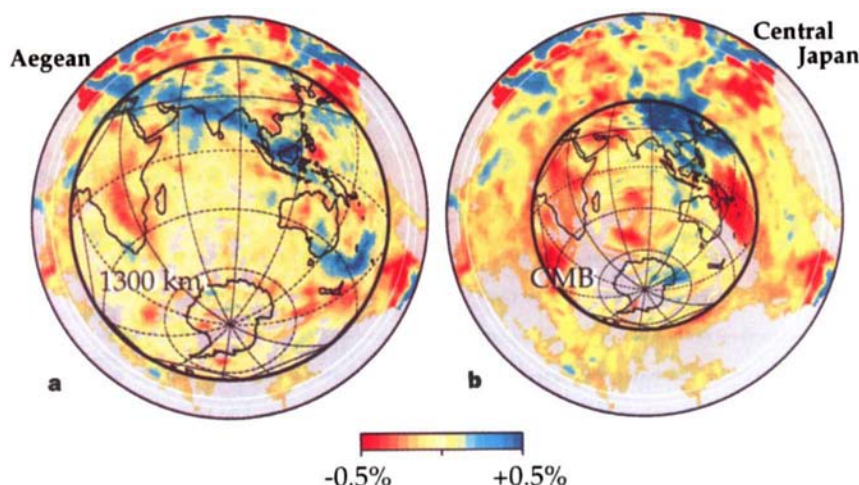


Figure 6 Mantle cross-sections along the great circle depicted in Fig. 1a to a depth of 1,300 km (a) and 2,750 km (b). In the centre of the vertical sections we plotted in map view the anomalies at the corresponding depth (that is 1,300 and 2,750 km) in an equal-area projection. The cross-section through the upper 1,300 km of the mantle demonstrates that the deep subduction of the African plate beneath eastern Europe²³ forms the western edge of the long linear structure that continues as an intermittent high-wave-speed anomaly beneath the southern margin of Asia and connects to the deep slab beneath Indonesia²⁵. Beneath eastern Europe, the Aegean slab does not seem to have sunk deeper than about 1,500 km, in accord with the narrow width of the western part of the Mesozoic Tethys ocean. The depth to the leading edge of this slab is well constrained.

Beneath central Japan we infer that the slab plunges into the lower mantle with an increased dip angle, which corroborates earlier studies based on residual sphere analyses⁸ and travel-time inversion¹⁹. Beneath central Japan the narrow structure seems to be continuous to the core–mantle boundary, where it connects to the well studied seismic anomaly beneath southeastern Asia. This vertical continuity is only observed for rather narrow segments of the Japan slab; the results of extensive resolution tests suggest, however, that such detail can be resolved by the data used (for instance Fig. 2c, d). The map view at 1,300 km reveals that this deep subduction of the Pacific plate has not resulted in a linear structure large enough to be detected by our current study. This lack of structure is representative for the depth range between 800 and 1,400 km.

CMB. We argue that this structural complexity is in keeping with the complex tectonic evolution of the region and that mantle stratification is a local and transient^{30–32} phenomenon with a characteristic timescale that is much shorter than that of mantle-wide overturn⁴⁷.

Lithospheric plates had been subducting along the eastern margin of continental Asia since at least 140 Myr ago (ref. 8) when Eocene (~45 Myr ago) plate reorganizations in the Pacific realm caused the separation of Japan from continental Asia and initiation of subduction beneath the Philippine Sea plate along the proto Izu Bonin and Mariana trenches. Post-Eocene clockwise migration of the newly formed arc system resulted in slab deflection in the transition zone beneath the Izu Bonin arc⁴⁷, which switched off the supply to the older downwelling in the lower mantle beneath the east Asian margin. The formation of the inferred 600 km or so slab window in the uppermost lower mantle during the past 40 Myr implies a minimum sinking rate of about 1.5 cm yr^{-1} in the lower mantle, which concurs with inferences from South America²². Trench migration may have caused local deflection of the Japan slab⁴⁷, but beneath central Japan subduction of the Pacific plate seems to have continued along a shallowly dipping conduit to a previously established downwelling anchored in the lower mantle⁴⁸ (Fig. 6b). Horizontal slabs are gravitationally unstable⁴⁹ and will eventually become entrained in lower-mantle flow through Rayleigh–Taylor instabilities; this process may produce narrow downwellings that may be hard to detect by seismic imaging.

Is there a lower-mantle transition zone? We infer from the images that flow in the shallow mantle connects, at least locally, to structural heterogeneity just above the CMB (Figs 5 and 6b), confirming previous suggestions of such a relationship^{12,13,27,43,50}, but the long narrow structures in the mid-mantle are strikingly different from the long-wavelength features just above the CMB (compare, for instance, Fig. 1d and f). This suggests that the downwellings that reach the CMB (Fig. 6b) lose the characteristic planar geometry across a transitional interval (approximately

1,800–2,300 km depth). A change in shape of the downwellings from sheet-like to more cylindrical is in accord with numerical flow simulations⁵¹. Comparison with other models also suggests that the change in character of heterogeneity is real. We remark that the inferred transition in heterogeneity coincides with a change in proportionality between P and S speed⁵², indicated by the Poisson's ratio, which may indicate a compositional change⁵³. There are several caveats. Although resolution tests demonstrate that the inferred breakdown of the linear structures is not a result of reduced sampling (Fig. 4) we realize that data coverage in the lowermost mantle is not satisfactory. Moreover, structural signal from heterogeneity in the deepest mantle may be lost from travel-time data owing to a mechanism known as wavefront healing⁵⁴. At this stage we cannot exclude the possibility that the change in geometry of the anomalies is causally related to past changes in plate motion, without the need for a change in physical state. Conversely, it is exciting to realize that the detailed structures revealed by global imaging may now begin to constrain reconstructions of past plate motion.

It is obvious that several aspects of the tomographic model presented can—and will—be improved. There are important issues that require further study (uneven sampling, interpretation of deep structure, remaining differences with long-wavelength models), but there seems to be increasing consensus from seismology for deep convective circulation in the mantle. A concerted effort of seismic imaging, plate reconstruction, numerical flow modelling, and chemical mass-balancing is required to unravel remaining mysteries of the deep Earth and to explain the flow behaviour in a lower-mantle transition zone and the isotope record. The increasing similarity between flow patterns deduced from global seismic imaging and numerical flow simulation^{55,56} (H.-P. Bunge and M. Richards, personal communication), and the fact that, for the first time, independent global wave-speed models begin to show excellent correlation of small-scale (<1,000 km) structure,⁴² mark

important and exciting contributions towards understanding the dynamic and thermal evolution of our planet. □

Received 2 September 1996; accepted 11 March 1997.

1. Dzierwowski, A. M., Hager, B. H. & O'Connell, R. J. Large-scale heterogeneities in the lower mantle. *J. Geophys. Res.* **82**, 239–255 (1977).
2. Dzierwowski, A. M. Mapping the lower mantle: determination of lateral heterogeneity in P velocity up to degree and order 6. *J. Geophys. Res.* **89**, 5929–5952 (1984).
3. Carlson, R. W. Mechanisms of Earth differentiation: consequences for the chemical structure of the mantle. *Rev. Geophys.* **32**, 337–362 (1994).
4. Hofmann, A. Mantle geochemistry: the message from oceanic volcanism. *Nature* **385**, 219–229 (1997).
5. Hager, B. H., Clayton, R. W., Richards, M. A., Comer, R. P. & Dzierwowski, A. M. Lower mantle heterogeneity, dynamic topography and the geoid. *Nature* **313**, 541–545 (1985).
6. Montagner, J.-P. Can seismology tell us anything about convection in the mantle? *Rev. Geophys.* **32**, 115–138 (1994).
7. Ritzwoller, M. H. & Lavelle, E. M. Three-dimensional seismic model of the Earth's mantle. *Rev. Geophys.* **33**, 1–66 (1995).
8. Richards, M. A. & Engebretson, D. C. Large-scale mantle convection and the history of subduction. *Nature* **355**, 437–440 (1992).
9. Ricard, Y., Richards, M. A., Lithgow-Bertelloni, C. & Le Stunff, Y. A geodynamic model of mantle density heterogeneity. *J. Geophys. Res.* **98**, 21895–21909 (1993).
10. Silver, P. G., Carlson, R. W. & Olson, P. Deep slabs, geochemical heterogeneity, and the large-scale structure of mantle convection: investigations of an enduring paradox. *Rev. Earth Planet. Sci.* **16**, 477–541 (1988).
11. Jordan, T. H., Lerner-Lam, A. L. & Creager, K. C. in *Mantle Convection* (ed. Peltier, W. R.) 98–201 (Gordon & Breach Scientific, New York, 1989).
12. Davies, G. F. & Richards, M. A. Mantle convection. *J. Geol.* **100**, 151–206 (1992).
13. Lay, T. The fate of descending slabs. *Annu. Rev. Earth Planet. Sci.* **22**, 33–61 (1994).
14. Shearer, P. & Masters, G. Global mapping of topography on the 660-km discontinuity. *Nature* **355**, 791–796 (1992).
15. Ito, E. & Takahashi, E. Postspinel transformations in the system Mg₂SiO₄–Fe₂SiO₄ and some geophysical implications. *J. Geophys. Res.* **94**, 10637–10646 (1989).
16. Christensen, U. R. & Yuen, D. A. The interaction of subducting lithospheric slab with a chemical or phase boundary. *J. Geophys. Res.* **89**, 4389–4402 (1984).
17. Puster, P. & Jordan, T. H. How stratified is mantle convection? *J. Geophys. Res.* **102**, 7625–7646 (1997).
18. Creager, K. C. & Jordan, T. H. Slab penetration into the lower mantle below the Mariana and other island arcs of the northwest Pacific. *J. Geophys. Res.* **91**, 3573–3589 (1986).
19. Kamiya, S., Miyatake, T. & Hirahara, K. How deep can we see the high velocity anomalies beneath the Japan island arcs? *Geophys. Res. Lett.* **15**, 828–831 (1988).
20. Van der Hilst, R. D., Engdahl, E. R., Spakman, W. & Nolet, G. Tomographic imaging of subducted lithosphere below northwest Pacific island arcs. *Nature* **353**, 37–43 (1991).
21. Fukao, Y., Obayashi, M., Inoue, H. & Nenbai, M. Subducting slabs stagnant in the mantle transition zone. *J. Geophys. Res.* **97**, 4809–4822 (1992).
22. Grand, S. P. Mantle shear structure beneath the Americas and the surrounding oceans. *J. Geophys. Res.* **99**, 11591–11621 (1994).
23. Spakman, W., Van der Lee, S., Van der Hilst, R. D. Travel-time tomography of the European-Mediterranean mantle down to 1400 km. *Phys. Earth Planet. Inter.* **79**, 3–74 (1993).
24. Van der Hilst, R. D. Complex morphology of subducted lithosphere in the mantle beneath the Tonga trench. *Nature* **374**, 154–157 (1995).
25. Widiyantoro, S. & Van der Hilst, R. D. The slab of subducted lithosphere beneath the Sunda arc, Indonesia. *Science* **271**, 1566–1570 (1996).
26. King, S. D. The viscosity structure of the mantle. *Rev. Geophys. Suppl.* **33**, 11–17 (1995).
27. Loper, D. E. A simple model of whole mantle convection. *J. Geophys. Res.* **90**, 1809–1836 (1985).
28. Hedlin, M., Shearer, P. & Earle, P. S. Seismic evidence for small-scale heterogeneity throughout the Earth's mantle. *Nature* (in press).
29. Krüger, F., Weber, M., Scherbaum, F. & Schlittenhardt, J. Evidence for normal and inhomogeneous lowermost mantle and core-mantle boundary structure under the Arctic and northern Canada. *Geophys. J. Int.* **122**, 637–657 (1995).
30. Thoraval, C., Machetel, P. & Cazenave, A. Locally layered convection inferred from dynamic models of the Earth's mantle. *Nature* **375**, 777–780 (1995).
31. Machetel, P. & Weber, P. Intermittent layered convection in a model mantle with an endothermic phase change at 670 km. *Nature* **350**, 55–57 (1991).
32. Tackley, P. Mantle dynamics: Influence of the transition zone. *Rev. Geophys. Suppl.* **33**, 275–282 (1995).
33. Su, W.-J., Woodward, R. L. & Dzierwowski, A. M. Degree 12 model of shear velocity heterogeneity in the mantle. *J. Geophys. Res.* **99**, 6945–6981 (1994).
34. Li, X.-D. & Romanowicz, B. Global mantle shear-velocity model developed using nonlinear asymptotic coupling theory. *J. Geophys. Res.* **101**, 22245–22272 (1996).
35. Masters, G., Johnson, S., Laske, G. & Bolton, H. A shear-velocity model of the mantle. *Phil. Trans. R. Soc. Lond. A* **354**, 1385–1411 (1996).
36. Masters, G., Jordan, T. H., Silver, P. G. & Gilbert, F. A spherical Earth structure from fundamental spheroidal-mode data. *Nature* **298**, 609–613 (1982).
37. Inoue, H., Fukao, Y., Tanabe, K. & Ogata, Y. Whole mantle P-wave travel-time tomography. *Phys. Earth Planet. Inter.* **59**, 294–328 (1990).
38. Vasco, D. W., Johnson, L. R., Pulliam, R. J. & Earle, P. S. Robust inversion of IASP91 travel time residuals for mantle P and S velocity structure. *J. Geophys. Res.* **99**, 13727–13755 (1994).
39. Engdahl, E. R., Van der Hilst, R. D. & Buland, R. P. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. *Bull. Seismol. Soc. Am.* (submitted).
40. Kennett, B. L. N., Engdahl, E. R. & Buland, R. Constraints on seismic velocities in the Earth from traveltimes. *Geophys. J. Int.* **122**, 108–124 (1995).
41. Nolet, G. Solving or resolving inadequate and noisy tomographic systems. *J. Comput. Phys.* **61**, 463–482 (1985).
42. Grand, S. P., Van der Hilst, R. D. & Widiyantoro, S. Global seismic tomography: a snapshot of convection in the earth. *Geol. Soc. Am. Today* (in the press).
43. Wyssession, M. E. Continents of the core. *Nature* **381**, 373–374 (1996).
44. Lithgow-Bertelloni, C., Richards, M. A., Ricard, Y., O'Connell, R. J. & Engebretson, D. C. Toroidal-poleoidal partitioning of plate motions since 120 Ma. *Geophys. Res. Lett.* **20**, 375–378 (1993).
45. Jordan, T. H. & Lynn, W. S. A velocity anomaly in the lower mantle. *J. Geophys. Res.* **79**, 2679–2685 (1974).
46. Wen, L. & Anderson, D. L. The fate of slabs inferred from seismic tomography and 130 million years of subduction. *Earth Planet. Sci. Lett.* **133**, 185–198 (1995).
47. Van der Hilst, R. D. & Seno, T. Effects of relative plate motion on the deep structure and penetration depth of slabs below the Izu-Bonin and Mariana island arcs. *Earth Planet. Sci. Lett.* **120**, 375–407 (1993).
48. Griffiths, R. W., Hackney, R. & Van der Hilst, R. D. A laboratory investigation of trench migration and the fate of subducted slabs. *Earth Planet. Sci. Lett.* **133**, 1–17 (1995).
49. Christensen, U. R. The influence of trench migration on slab penetration into the lower mantle. *Earth Planet. Sci. Lett.* **140**, 27–39 (1996).
50. Kendall, J.-M. & Silver, P. G. Constraints from seismic anisotropy on the nature of the lowermost mantle. *Nature* **381**, 409–412 (1996).
51. Bercovici, D., Schubert, G. & Glatzmaier, G. 3-dimensional spherical-models of convection in the Earth's mantle. *Science* **244**, 950–955 (1989).
52. Bolton, H. & Masters, G. A region of anomalous $d \ln V_p / d \ln V_p$ in the deep mantle. (abstr.) *Eos* **77**, F687 (1996).
53. Loper, D. & Lay, T. The core-mantle boundary region. *J. Geophys. Res.* **100**, 6397–6420 (1995).
54. Nolet, G. & Moser, T.-J. Teleseismic delay times in a 3-dimensional Earth and a new look at the S-discrepancy. *Geophys. J. Int.* **114**, 185–195 (1993).
55. Puster, P., Hager, B. H. & Jordan, T. H. Mantle convection experiments with evolving plates. *Geophys. Res. Lett.* **22**, 2223–2226 (1995).
56. Bunge, H.-P., Richards, M. A. & Baumgardner, J. R. The effect of viscosity stratification on mantle convection. *Nature* **379**, 436–438 (1996).

Acknowledgements. We thank B. Kennett and J. Braun for providing graphics software; S. Grand, C. Lithgow-Bertelloni and W.-J. Su for access to their models; T. Jordan, B. Hager, S. Grand and C. Froidevaux for discussions; and G. Masters for a review. This work was supported in part by the US National Science Foundation.

Correspondence and requests for materials should be addressed to R.D.v.d.H. (e-mail: hilst@mit.edu).

Detection of the Earth's rotation using superfluid phase coherence

Keith Schwab, Niels Bruckner & Richard E. Packard

Physics Department, University of California, Berkeley, California 94707, USA

It has long been recognized that the macroscopic quantum properties of superfluid helium could form the basis of a technique for measuring the state of absolute rotation of the containment vessel^{1–5}: circulation of superfluid helium is quantized, so providing a reference state of zero rotation with respect to inertial space. Here we provide experimental proof of this concept by detecting the rotation of the Earth using the spatial phase coherence of superfluid ⁴He, thus providing independent corroboration of an earlier report⁶ that demonstrated the feasibility of making such a measurement. Our superfluid container is constructed on a centimetre-size silicon wafer, and has an essentially toroidal geometry but with the flow path interrupted by partition incorporating a sub-micrometre aperture. Rotation of the container induces a measurable flow velocity through the aperture in order to maintain coherence in the quantum phase of the superfluid. Using this device, we determine the Earth's rotation rate to a precision of 0.5% with a measurement time of one hour, and argue that improvements in sensitivity of several orders of magnitude should be feasible.

The idea at the foundation of the experiment is that the superfluid state is described by a macroscopic quantum phase, ϕ , whose gradient is proportional to the superfluid velocity. This leads to the conclusion that, if the wavefunction is single-valued, the integral of the velocity around a closed curve (that is, the circulation) must be quantized:

$$\oint \mathbf{v}_s \cdot d\mathbf{l} = \frac{h}{2\pi m} \oint \nabla \phi \cdot d\mathbf{l} = n \frac{h}{m} \quad n = 0, 1, 2, \dots \quad (1)$$

Here, h is Planck's constant and m is the atomic mass of ⁴He.

In a container rotating at small angular velocity, the ground state of the superfluid has zero circulation. If the container is not symmetric with respect to the rotation axis, the superfluid must move to avoid the moving boundaries, but the induced flow must always satisfy equation (1).

We now consider a toroidal container of average radius R , interrupted by a septum which is pierced by a submicrometre aperture (Fig. 1a); the torus is slowly rotating at angular velocity ω . The moving septum induces the fluid in the body of the torus to flow, almost like a solid body ($v_s = \omega R$). To satisfy equation (1), when the circulation integral traverses the torus and passes through the aperture, there must be an induced velocity v_a in the aperture, that cancels the solid-body flow contributions from the bulk of the torus. The rotation-induced velocity in the aperture is given by

$$v_a = \frac{2\omega\pi R^2}{l_{\text{eff}}} \cos \theta \quad (2)$$

where R is the radius of the torus and l_{eff} is an effective length of the aperture. The $\cos \theta$ factor accounts for the general case when the axis of rotation is inclined at an angle θ with respect to the normal to the toroidal plane.

Our experiment consists of measuring the rotation-induced velocity with apparatus which is a superfluid analogue of a superconducting a.c. SQUID^{7,8}.

The device, shown in Fig. 1b, is a Helmholtz oscillator similar to that used to detect single phase slips⁹. A description of the device

is given in the figure legend. The tension of the membrane that seals the upper surface of the device provides the potential energy reservoir, and the fluid passing through the central microaperture and around the peripheral channel provides the kinetic-energy reservoir. The oscillator has a resonant frequency of 66 Hz and exhibits a Q of 20,000 at $T = 280$ mK, the typical operating temperature.

Figure 2 is a plot of the oscillation amplitude as a function of the applied electric drive. The observed staircase pattern is a characteristic feature of these double-hole devices¹⁰. The plateau represented by the first step is due to the fluid reaching the intrinsic critical velocity for quantized vortex creation in the aperture. When a vortex is created in the hole, it passes across the aperture removing a discrete amount of energy from the flow. The quantum phase

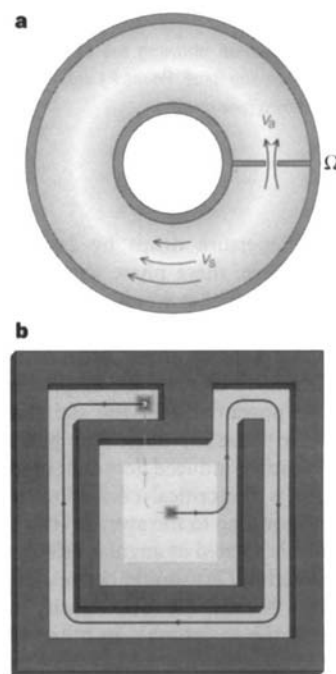


Figure 1 **a**, A torus rotating at angular velocity Ω , partitioned by a septum containing a small aperture. Backflow through the aperture, at velocity v_a , is induced by the rotation to keep the circulation equal to zero. The superfluid velocity v_s approximates that of a solid body. **b**, The superfluid rotation sensor built on a silicon wafer¹¹. A central square depression continues around the periphery of the chip and terminates near a relatively large window (1 mm square) opening to the back of the wafer. The depressed areas (80 μm deep) are covered and sealed with an 8- μm -thick Kapton diaphragm epoxied to the raised silicon surface. A single small aperture¹² (190 nm $\times$ 1,000 nm in a 100-nm-thick silicon nitride wall), also opening to the back of the wafer, is placed in the centre of the central square depression. The device is placed in a superfluid-filled enclosure. There is a continuous path within the fluid: from the top surface of the microaperture, around the peripheral channel, through the large hole to the back of the wafer, and through the microaperture back into the central depression. This path serves as the toroidal geometry mentioned above. The mean length, L , of each side of the channel is 10.5 mm and the channel width is 1.5 mm. If the device is rotated, the induced velocity in the aperture is given by equation (2) with the effective area (πR^2 in equation (2)) equal to $0.93L^2$. The arrowed line indicates a path of integration for the circulation. The solid segment lies within the etched channels and the dotted line represents a path on the back surface of the wafer. The Kapton membrane is metallized on both sides, and can be moved by applying an electric potential to an electrode positioned in the central depression. The membrane's resultant position is monitored with a superconducting displacement transducer¹³ having a resolution of 10^{-15} m Hz^{-1/2}.

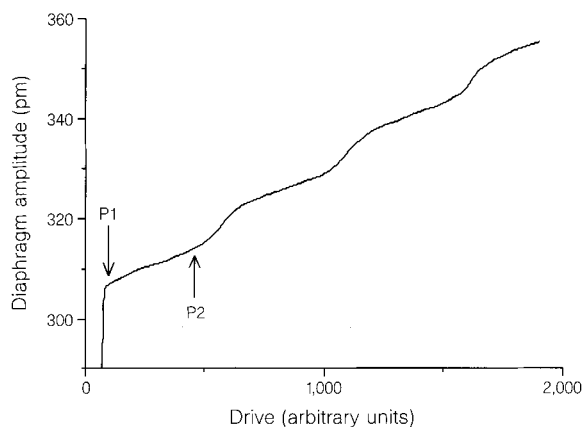


Figure 2 The staircase response showing the oscillation amplitude of the diaphragm as a function of driving force. Points P1 and P2 are two typical points monitored during reorientation (see text).

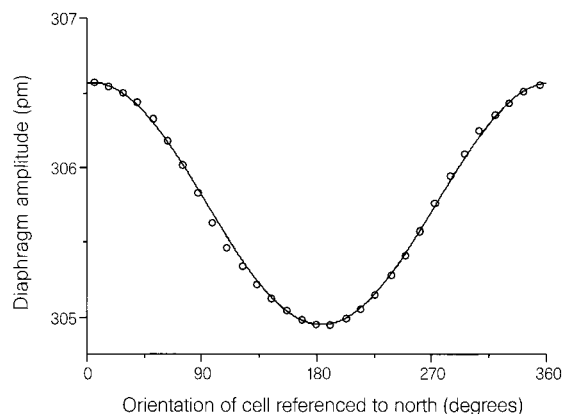


Figure 3 The oscillation amplitude at point P1 (in Fig. 2) as a function of orientation of the chip with respect to north. The oscillation amplitude is smoothed by a lock-in amplifier with a time constant of 10 s. The data (circles), which are digitally recorded at 0.5 points per s, is then averaged with a sliding average using 250 points. Point P2 shows a similar modulation but is out of phase by π . The solid curve is a fit of the data to a cosine function.

difference across the aperture drops by 2π . As the oscillator proceeds through its cycle, these phase slips occur in pairs and limit the amplitude increase that would otherwise accompany an increasing drive level. Subsequent steps are produced when, after an even number of phase slips in every oscillation half-cycle, the driving force is sufficient to increase the amplitude until addition slip pairs occur. The upward tilt in the staircase is due to the stochastic nature of the phase slip nucleation process⁹.

In the absence of rotation-induced flow and trapped vorticity, the level of the first step is the critical oscillation amplitude of the membrane, A_c , corresponding to the average critical velocity in the aperture. If the device is rotated at angular velocity ω , the induced velocity (given by equation (2)) should create an apparent shift in the level of the plateau because the total velocity in the aperture is not directly sensed by the membrane. In other words, the membrane motion is responsible for only a fraction of the actual velocity in the aperture. The remaining fraction includes the rotation-induced flow plus residual bias flow from trapped vorticity. Thus the oscillation amplitude on the n th plateau A_n is given by

$$A_n = A_c - B\omega + C \quad (3)$$

where B is a known geometric factor involving the areas of the membrane and the aperture and C represents the effect of bias currents. On a given plateau, A_n is bounded between levels corresponding to 0 and π phase difference across the aperture. Therefore an unbounded increase in ω would result in a triangle response in A_n .

The normal to the plane of our torus (a silicon wafer) lies in a horizontal plane. The device can be reoriented with respect to the compass by rotating the entire cryostat about a vertical axis. If one monitors the level on a given point of a plateau, we expect the level to shift with rotation flux. Our sensor has sufficient loop area that the rotation of the Earth at our latitude (38° N) corresponds to a maximum phase shift of $\sim \pm \pi/4$. If the normal to the plane of the torus points in an east–west direction, the Earth's rotation creates no flux through the loop. If the normal is oriented at angle α with respect to a meridian, the Earth's rotation flux through the loop is $2\omega\pi R^2 \cos 38^\circ \cos \alpha$. Thus when the apparatus is reoriented around a vertical axis we expect a $\cos \alpha$ dependence of the amplitude of point P1 on the first plateau in Fig. 2. Other points along the staircase are expected to be modulated with α in a periodic manner, possibly offset and inverted. The expected modulation pattern depends on the value of any fixed bias phase difference that might

exist across the aperture. These phase biases may arise from vortex lines pinned anywhere in the flow path. The motion of a single vortex across the peripheral channel causes a 2π shift across the aperture. Therefore the successful operation of the rotation sensor depends on there being no vortex motion during the course of a measurement. In practice this is difficult to achieve. The data must be recorded at night, when mechanical noise arising from human activity in the building is at a minimum.

Figure 3 shows the measured correlation between point P1 (Fig. 2) on the first plateau level and the orientation angle of the torus, α . To record this data the entire cryostat is smoothly rotated at about one revolution per hour.

The solid line drawn through the data is a fit to a cosine function. To predict the expected peak-to-peak modulation depth it is necessary to identify the extreme levels of A_n corresponding to 0 and π phase difference across the aperture. Although we never know the exact bias state, we can make a close estimate of the expected modulation, based on a staircase that has the features of a 0 bias case (equal length steps). This method will slightly overestimate the depth of the modulation. The observed depth of the modulation is 1.6×10^{-12} m which, based on the above mentioned procedure, corresponds to a phase difference across the aperture of 0.51π . Based on the computed flow pattern in the device (we use a commercial program to solve Laplace's equation in the channel) we predict that the rotation of the Earth will make a phase difference across the aperture of 0.46π . The small difference between the observed number and predicted value is due to the systematic uncertainties in determining the zero bias staircase and in predicting the flow pattern in the device. Random uncertainty arises from the stochastic nature of the phase slip nucleation mechanism and is understood in the context of a thermal activation model.⁷

We have also observed that the modulation of other points on the staircase follows equation (3). For instance, point P2 in Fig. 2 is found to change as $-\cos \alpha$, which is the expected response.

When the bias is anywhere between approximately $\pi/4$ and $3\pi/4$, the modulation of the first plateau should pass through $\cos \alpha = 0$ when the normal to the toroidal plane is oriented in an east–west direction. We find the median level in the modulation to occur when the toroidal plane normal points within $\sim 5^\circ$ of true east–west. The uncertainty is systematic and is due to a lack of knowledge of the precise orientation of the silicon wafer within the cryostat.

It is interesting to consider the projected sensitivity of this type of device for terrestrial rotation measurements. The present device has

the sensitivity to detect a change in the Earth's rotation, Ω_E , of $\sim 0.5\%$ in one hour. The peripheral channel could be made with multiple turns and could also be enlarged in scale. For a 30-turn tubular channel, with $R = 3$ cm and channel diameter of 1.6 mm, a one-hour measurement time could resolve a rotation change equal to $5 \times 10^{-6} \Omega_E$. To achieve this projected sensitivity, vorticity in the tube must be completely pinned, for instance by filling the tube with a porous medium. It remains to be seen if this superfluid rotation sensor would surpass other techniques such as ring laser gyroscopes, atomic interferometers or spinning balls. $\square$

Received 12 December 1996; accepted 10 March 1997.

- Gregory, D. D. Observations of low-lying superfluid states in a rectangular annulus. Thesis, Univ. California, San Diego (1972).
- Guernsey, R. W. in *Proc. 12th Int. Conf. on Low Temperature Physics* 79–81 (Keigaku, Tokyo, 1971).
- Crooker, B. C. Studies of the flow properties of superfluid ^4He and of the superfluid density of ^4He films in restricted geometries. Thesis, Cornell Univ. (1984).
- Anandan, J. The Josephson effect in superfluid helium and general relativity. *J. Phys. A* **17**, 1367–1380 (1984).
- Bonaldi, M., Vitale, S. & Cerdonio, M. Rotationally induced dissipation in superfluid helium. *Phys. Rev. B* **42**, 9865–9874 (1990).
- Avenel, O. & Varoquaux, E. Detection of the Earth rotation with a superfluid double-hole resonator. *Czech. J. Phys. (Suppl. S6)* **48**, 3319–3320 (1996).
- Packard, R. E. & Vitale, S. Principles of superfluid helium gyroscopes. *Phys. Rev. B* **46**, 3540–3549 (1992).
- Schwab, K. C. Experiments with superfluid oscillators: design and microfabrication of a superfluid gyroscope; modulation of a ^4He rf SQUID analog by the Earth's rotation. Thesis, Univ. California, Berkeley (1996).
- Zimmermann, Wm. The flow of superfluid ^4He through submicron apertures: phase slip and critical velocities due to quantum vortex motion. *Contemp. Phys.* **37**, 219–234 (1996).
- Avenel, O. & Varoquaux, E. Josephson Effect and quantum phase slippage in superfluids. *Phys. Rev. Lett.* **60**, 416–419 (1988).
- Schwab, K., Steinhauer, J. Davis, J. C. & Packard, R. E. Fabrication of a silicon-based superfluid oscillator. *J. Microelectromech. Syst.* **5**, 180–186 (1996).
- Amar, A., Lozes, R., Sasaki, Y., Davis, J. C. & Packard, R. E. Fabrication of submicron apertures in thin membranes of silicon nitride. *J. Vac. Sci. Technol. B* **11**, 259–262 (1993).
- Paik, H. J. Superconducting tunable-diaphragm transducer for sensitive acceleration measurements. *J. Appl. Phys.* **47**, 1168–1178 (1976).

Acknowledgements. We thank A. Amar, J. C. Davis, Yu. Mukharsky and J. Steinhauer for help with earlier versions of this experiment; A. Loshak for help with fabrication of the silicon device; S. Vitale, E. Varoquaux, O. Avenel and S. Backhaus for conversations; and R. Orr for preparing Fig. 1. This work was supported in part by the Air Force Office of Scientific Research, the Office of Naval Research, and the National Science Foundation.

Correspondence and requests for materials should be addressed to R.E.P. (e-mail: packard@garnet.berkeley.edu).

Non-volatile memory device based on mobile protons in SiO_2 thin films

K. Vanheusden*, W. L. Warren*, R. A. B. Devine†, D. M. Fleetwood*, J. R. Schwank*, M. R. Shaneyfelt*, P. S. Winokur* & Z. J. Lemnios‡

* Sandia National Laboratories, Albuquerque, New Mexico 87185-1349, USA

† France Telecom/CNET, BP 98, Meylan Cedex, France

‡ Defense Advanced Research Projects Agency, Arlington, Virginia 22203-1714, USA

The silicon/silicon-dioxide system provides the cornerstone of integrated-circuit technology¹. Since the introduction of devices based on this system, the (largely deleterious) effects on device operation of mobile and trapped charges in the oxide layer have been studied in great detail. Contamination by alkali ions, for example, was a major concern in the early days of metal-oxide-semiconductor device fabrication². But not all SiO_2 impurities are undesirable: the addition of hydrogen, for example, has the beneficial property of rendering charge traps inactive¹. Here we show that mobile H^+ ions introduced by annealing into the buried oxide layer of $\text{Si}/\text{SiO}_2/\text{Si}$ structures, rather than being detrimental, can form the basis of a non-volatile memory device. These mobile protons are confined to the oxide layer, and their space-charge

distribution can be controlled and rapidly rearranged at room temperature by an applied electric field. Memory devices based on this effect are expected to be competitive with current state-of-the-art Si-based memories, with the additional advantage of simplicity—only a few standard processing steps are required.

A variety of $\text{Si}/\text{SiO}_2/\text{Si}$ materials was investigated. A first group, silicon-on-insulator (SOI) materials, consisted of amorphous SiO_2 buried underneath a single crystalline thin Si film. This group included 'separation by the implantation of oxygen' (SIMOX) material formed by high-energy/high-dose ion implantation of oxygen into a Si wafer, and Unibond, which uses a thermally oxidized wafer and wafer-bonding technology to obtain the SOI structure. Apart from these 'technologically advanced' SOI materials, we also studied a second group of standard thermally grown SiO_2 capped with a poly-Si layer. The common processing treatment for all the above structures was a high-temperature 'formation anneal' (1,100–1,300 °C) in an inert atmosphere (1% O_2 added). It was observed that this anneal step was crucial as a pretreatment for the subsequent incorporation of mobile protons in the buried oxide, probably because it created defects in the buried oxide which may be catalyst sites for the generation of atomic hydrogen³. A schematic cross-section of the final structure can be seen in Fig. 1b. The top Si layer thickness was typically 200 nm, while the buried oxide varied between 1 μm and 40 nm.

Small isolated islands (3 mm²) of top Si layer were created to facilitate lateral diffusion of hydrogen (or deuterium) into the buried oxide during anneal treatments in forming gas (N_2 with 5 vol.% of H_2 or D_2 , 99.999% pure) and nitrogen (N_2 99.999% pure). Annealing in forming gas between 500 and 800 °C was found to induce a reversible flat-band voltage shift at the top-Si/ SiO_2 interface. Figure 2 shows this reversible shift in the drain-current/gate-voltage I_D – V_G curves measured on a SIMOX substrate. Similar features were observed in other SOI materials and in poly-Si capped thermal oxide substrates. No reversible shift could be observed after annealing in N_2 as opposed to forming gas, showing that the presence of hydrogen in the anneal atmosphere is crucial. A detailed

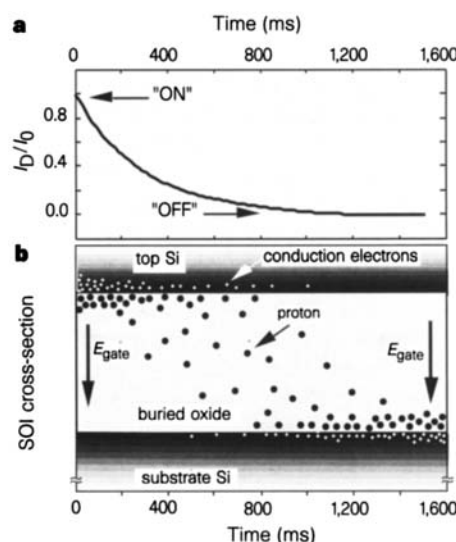


Figure 1 a. Normalized decay of drain current (I_D) versus time at room temperature in a point-contact device with a 40-nm layer of buried silicon oxide, annealed in forming gas at 600 °C. The positive ions were first drifted to the top-Si/ SiO_2 interface under positive gate bias (V_G) followed by switching to $V_G = -2$ V (-0.5 MV cm⁻¹). **b.** Cross-sectional schematic of the evolution of mobile charged species in the device on the same timescale, visualizing the applied electric field (E_{gate}) induced drift of the H^+ ions in the buried oxide from the top to the bottom interface, and its effect on the conduction electron density in the conduction channels near the interfaces.

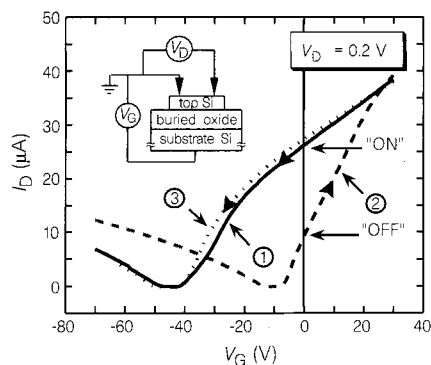


Figure 2 Drain-current/gate-voltage (I_D - V_G) curves measured on SIMOX (400-nm-thick buried oxide) after it received a 550°C forming-gas anneal. The curves were measured using the point-contact transistor method¹⁰. This simple test 'device' is based on the upside-down metal-oxide-semiconductor (MOS) structure using the specific SOI configuration as shown in the inset: the buried oxide plays the role of the gate oxide and the top Si layer represents the transistor body. Two metal probe tips are placed on the top Si layer to form the source and drain point contacts ($V_D = 0.2$ V), while the gate voltage (V_G) is applied to the back of the Si substrate. Curve 1 was recorded sweeping the gate bias from positive to negative. Subsequently, curve 2 was recorded using the opposite gate-voltage sweep direction (from negative to positive). Both curves were recorded at room temperature after keeping the gate at the initial bias value for 5 min. Curve 3 was recorded after curve 2 using the same procedure (with pre-bias) as for curve 1, showing the total reversibility of the process.

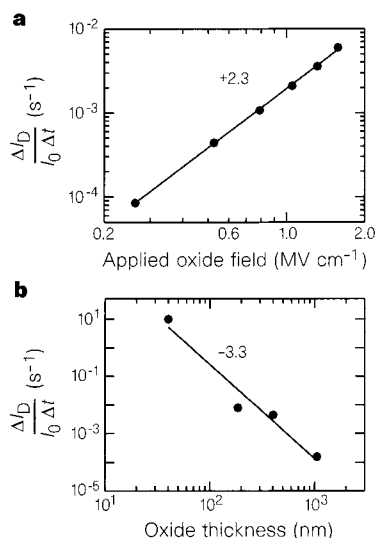


Figure 3 Initial rate of decay of the point-contact field-effect-transistor (FET) drain current obtained from I_D versus t curves such as the one shown in Fig. 1a. Devices were annealed in forming gas ($N_2 : H_2$, 95 : 5) at 600°C for 20 min. All data were recorded at room temperature by first drifting the ions to the top-Si/SiO₂ interface, followed by switching to a negative substrate bias. **a**, As a function of applied negative gate voltage for a fixed buried oxide thickness of 1 μm, that is, as a function of applied drift field. **b**, As a function of buried oxide thickness for a fixed applied gate voltage of -10 V. The solid lines are power-law ($y = ax^c$) least-square fits to the data; the resulting exponents are shown next to the lines.

capacitance/voltage and ionic out-diffusion study⁴, together with analysis of the drift kinetics as discussed below, provide substantial evidence that the observed reversible shift is due to mobile H⁺ ions, not to charge trapping or to other alkali ion contaminants² in the buried oxide.

In general, a volume charge distribution $\rho(x)$ in the oxide will result in a shift of the I_D - V_G curve along the voltage axis. This flat-band voltage shift (ΔV_{FB}) depends on the charge sign, the charge density, and its spatial distribution in the SiO₂ layer². For the top-Si, ΔV_{FB} is maximized if the protons are located in the oxide near the top-Si/SiO₂ interface and is minimized if the proton charge is located near the substrate-Si/SiO₂ interface. The observed reversible shift in the I_D - V_G curves is the result of an electric-field-induced migration of protons from one Si/SiO₂ interface to the other. The negative sign of the voltage shift ΔV_{FB} confirms that the migrating ions are positive, with an areal density of $\sim 2 \times 10^{12}$ cm⁻².

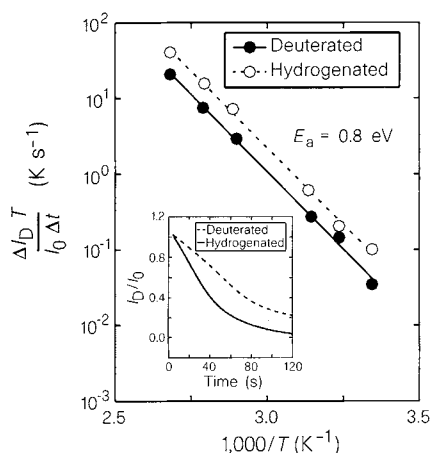
To provide additional evidence that the reversible voltage shift is actually due to mobile protons, and to investigate their kinetics in thin SiO₂ films, we analysed the time dependence of the charge migration as a function of V_G , oxide thickness (d), and temperature (T). By applying a positive bias V_G to the substrate of Si/SiO₂/Si device structures, the positive ions were accumulated in a thin boundary layer at the top-Si/SiO₂ interface. The resulting positive charge density in the oxide at the top-Si/SiO₂ interface, Q_{H^+} , induces an image charge of equal density and opposite sign at the interface in the top Si layer. As a result, a conduction channel is formed in the top Si layer at the top-Si/SiO₂ interface, and a drain current (I_D) will flow in the point-contact device. As a first-order approximation, a linear relationship can be derived between Q_{H^+} and the resulting

drain current I_D . Consequently, the decrease in Q_{H^+} as a function of time, $\Delta Q_{H^+}/\Delta t$, caused by switching the gate bias polarity which drifts the ions across the oxide layer towards the substrate interface, can be measured as a decrease in I_D over time ($\Delta I_D/\Delta t$). Figure 1 shows this decay of I_D versus time together with a schematic illustrating the redistribution of charged species.

The ionic transport from one interface to the other through the buried SiO₂ depends strongly on the ionic species involved; it is a complicated function of time, temperature, gate voltage and oxide thickness. A variety of formalisms and approximations have been proposed to describe this process adequately^{1,2}. To study the H⁺ drift kinetics in the buried oxide, we used the initial rate of decay of the drain current ($\Delta I_D/\Delta t$) obtained from I_D versus t curves such as the one shown in Fig. 1a as the characteristic parameter. In Fig. 3a we plot this initial slope against the applied negative oxide field, showing a square dependence. A near cubic dependence of the initial slope on the thickness of the buried oxide layer at a fixed gate voltage is derived from the plot in Fig. 3b. The data in Fig. 3 suggest that $\Delta Q_{H^+}/\Delta t$ is best described as space-charge-limited current³. Such dependence indicates charge transport (proton drift) through—rather than charge trapping and detrapping in—the buried oxide. Solving Poisson's equation for this problem yields:

$$\frac{\Delta Q_{H^+}}{\Delta t} = J_{H^+} \propto \frac{V_G^2}{d^3} \mu_{H^+} \quad (1)$$

where J_{H^+} is the space-charge-limited ionic current density through the SiO₂ layer, and μ_{H^+} is the ionic mobility. To first order, the ionic mobility in equation (1) is governed by a thermally activated hopping process (jumping from site to site)⁶.



To determine the thermal activation energy E_a for H^+ motion, the initial rate of decay of the drain current $\Delta I_D / \Delta t$ was measured as described above, while the temperature was varied at a fixed V_G . Figure 4 shows the Arrhenius plots. From these plots, we deduce an activation energy $E_a \approx 0.8$ eV for the ionic mobility in both the hydrogenated and deuterated buried SiO_2 . This value is in good agreement with the drift activation energy of H^+ in SiO_2 thin films, found by other workers⁷⁻⁹. The difference in absolute values between the two curves in Fig. 4 results from a difference in a prefactor of 1.5, again in good agreement with the theoretically expected difference of $(m_D/m_H)^{1/2} = 2^{1/2}$ (see equation in Fig. 4 legend, where m_D and m_H are the atomic mass of D and H, respectively).

Today, fast-access data storage in computer systems is dominated by dynamic and static random-access memory (RAM) semiconductor devices. They have, however, one major drawback compared to non-volatile memories: they lose all their information when power is removed. It is easy to see how the reversible flat-band voltage shift observed in this work could be utilized in a non-volatile field-effect-transistor (NVFET) memory device. An n-channel transistor can be changed to 'normally on' or 'normally off' by applying a positive or negative gate (substrate) bias which will drift the protons to the top-Si/ SiO_2 or substrate-Si/ SiO_2 interface, respectively. For a memory device this can be interpreted as writing the device to a bit state '1' or '0', respectively. To read the device, the zero-bias drain current I_0 is simply measured (high current then corresponds to logic state '1', low current to '0'), as visualized in Figs 1 and 2.

The most important requirement for a memory device is that it retains the stored information over a specified time and temperature range. Retention tests were performed by drifting the protons to the top-Si/ SiO_2 or substrate-Si/ SiO_2 interface, followed by heating the device to 200 °C at floating gate for extended times. No instabilities in the initial I_D - V_G characteristic could be observed for times up to 25 h, in contrast with the classical picture of instabilities associated with mobile ion contaminants in SiO_2 (ref. 2). Furthermore, fatigue experiments performed on these structures show that the device easily endures over 10^4 write-erase cycles without degradation. This means that the loss of protons is negligible, that is, they are imprisoned between the two encapsulating Si layers.

For memory devices, a short write time is desirable. It follows from Fig. 3 that the device speed will be proportional to V_G^2 and d^{-3} . For test devices fabricated using a poly-Si-capped 40-nm thermal oxide substrate, a write time of ~ 50 ms was attained (write/erase voltage $V_G = \pm 4$ V). Write times as fast as 1 ms can be extrapolated from the fitted curve in Fig. 3b for 10-nm thermal oxides. The latter experimental result shows that these NVFET memory devices can be fabricated using thermal oxides capped with a poly-Si layer, a standard procedure in silicon-based metal-oxide-semiconductor (Si-MOS) processing.

Figure 4 Arrhenius plot of the initial rate of decay of the normalized drain current measured on point-contact FET devices with a 1- μ m buried oxide after hydrogenation or deuteration at 600 °C in forming gas (N_2H_2 or $N_2:D_2$, 95:5). Each data point was obtained in the same way as in Fig. 3 by first drifting the ions to the top-Si/ SiO_2 interface, followed by switching to -10 V substrate bias at different substrate temperatures. The thermal activation energy E_a for ionic motion is obtained from the slope of the curves by using the expression for the ionic mobility: $\mu = \{ (q a^2 \nu) / (6 k_B T) \} \exp(-E_a / k_B T)$, where q is the ionic charge, a is the intersite hopping distance, k_B is Boltzmann's constant and T is the absolute temperature. The parameter ν is proportional to $1/m^3$, where m is the mass of the ionic species. The inset shows the normalized decay of I_0 versus time at 73 °C.

This NVFET device has potential advantages over state-of-the-art non-volatile memory technologies such as electrically erasable programmable read-only memory (EEPROM). Its speed, retention and lifetime performance are expected to be competitive with these existing technologies, but it is simple in design, requires fewer processing steps and can potentially operate at very low voltages. □

Received 21 August 1996; accepted 19 February 1997.

1. Sze, S. M. in *Physics of Semiconductor Devices* (Wiley, New York, 1981).
2. Snow, E. H., Grove, A. S., Deal, B. E. & Sah, C. T. Ion transport phenomena in insulating films. *J. Appl. Phys.* **36**, 1664-1673 (1965).
3. Devine, R. A. B., Leray, J.-L. & Margail, J. Ultraviolet-radiation induced defect creation in buried SiO_2 layers. *Appl. Phys. Lett.* **59**, 2275-2277 (1991).
4. Vanheusden, K., Warren, W. L. & Devine, R. A. B. Incorporation of mobile H^+ ions in SiO_2 thin films by thermal annealing. *J. Appl. Phys.* (submitted).
5. Jayant Baliga, B. in *Modern Power Devices* 148-149 (Wiley, New York, 1987).
6. Elliott, S. R. in *Physics of Amorphous Materials* 245-247 (Longman Scientific & Technical, Harlow, 1990).
7. Hofstein, S. R. Proton and sodium transport in SiO_2 films. *IEEE Trans. Elect. Dev.* **ED-14**, 749-759 (1967).
8. Winokur, P. S., Boesch, H. E., McGarrity, J. M. & McLean, F. B. Field- and time-dependent radiation effects at the SiO_2/Si interface of hardened MOS capacitors. *IEEE Trans. Nucl. Sci.* **NS-24**, 2113-2118 (1977).
9. McLean, F. B. A framework for understanding radiation induced interface states in SiO_2 MOS structures. *IEEE Trans. Nucl. Sci.* **NS-27**, 1651-1657 (1980).
10. Cristoloveanu, S. & Li, S. S. in *Electrical Characterization of Silicon-on-Insulator Materials and Devices* 104-111 (Kluwer Academic, Boston, MA, 1995).

Acknowledgements. We thank C. Keast for the thermal oxide samples. The part of this work performed at Sandia National Laboratories was supported by the US Department of Energy.

Correspondence and requests for materials should be addressed to K.V. (e-mail: karel@unm.edu).

Dynamics of glasses below the glass transition

B. Jérôme & J. Commandeur

FOM-Institute for Atomic and Molecular Physics, Kruislaan 407, 1098 SJ Amsterdam, The Netherlands

A wide variety of materials ranging from metals to polymers can solidify as glasses rather than crystals. The glass transition is associated with a slowing down of molecular motion: a liquid becomes a glass when structural relaxation no longer occurs on experimentally accessible timescales. Most of our current knowledge about collective molecular motion in glass-forming materials is based on observations of the supercooled liquid state above the glass transition^{1,2}. The lack of direct information about molecular dynamics in the glass state itself leaves room for conflicting models of the glass transition²⁻⁹. Here we show that, by taking advantage of confinement effects in thin films, molecular

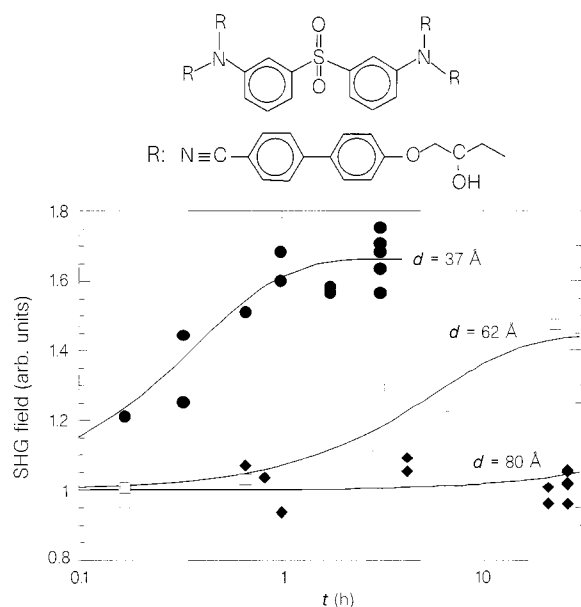


Figure 1 Time-dependence of the second-harmonic field generated by films of three different thicknesses at 323 K. The drawn curves are least-square fits to an exponential. Films of the glass-forming liquid crystal shown at the top (provided by Akzo-Nobel, Arnhem, The Netherlands²⁹) were prepared on float quartz plates by spin-coating from a solution in dimethylformamide. The different film thicknesses d were obtained by varying the concentration of the solution. The thickness of the films (found to remain constant with time) was determined by X-ray reflectivity measurements; all the reflectivity curves could be fitted by the reflectivity of an amorphous layer with the same density ($\pm 5\%$) for all thicknesses. The sudden evaporation of the solvent leaves behind a film that has not yet equilibrated. The relaxation of the molecules at the interface with the quartz plate is monitored by optical second-harmonic generation (SHG) measurements. The second-harmonic signal is almost exclusively due to the cyanobiphenyl liquid-crystal groups of the molecules in contact with the quartz substrate: they form a polar interfacial layer with the polar cyanogroup pointing towards the substrate²⁷. The rest of the film is centro-symmetric and generates a negligible contribution to the signal. At the free surface in particular, the liquid-crystal groups orient parallel to the surface, leading to a non-polar ordering (S. J. Picken and F. G. H. van Wijk, personal communication). The second-harmonic signal was measured using the 532-nm radiation from a frequency-double Q-switched mode-locked Nd:YAG laser as source. The signal was detected by photon counting. The incident beam (respectively second-harmonic signal) was polarized perpendicular (respectively parallel) to the plane of incidence. To avoid any effect of laser heating, the laser spot was moved between measurements.

dynamics can also be probed experimentally below the glass transition. We use second-harmonic generation to study the relaxation behaviour of molecules of a glass-forming liquid crystal confined in a thin film on a silica plate. Our measurements provide direct evidence that the collective character of molecular motion is responsible for the slowing down of mobility in glasses.

In many theories of the glass transition^{10–12}, it is assumed that the dynamics of glass-forming liquids is governed by the collective motion of molecules over a length scale ξ (the cooperative length) which increases with decreasing temperature; the observed slowing down of molecular motion is due to this increase of ξ . Although the concept of cooperative motion in glass-forming materials is quite old^{13–15}, an unambiguous test is still missing. Some measurements of the dynamics of supercooled liquids have been interpreted in terms of a cooperative length (see refs 11, 16 and references therein), but without proving the validity of this idea. Recently, qualitative indications of collective motion have been reported in simulations^{17–19} (but also contested²⁰), in studies of dependence of the relaxation of probe molecules on their size (ref. 11 and references therein), and in the observation of a decrease of the glass transition temperature of some glass-forming liquids in confined geometries^{21–25}.

Confinement allows one to explore different length scales in a more controlled way than by adding probe molecules of different sizes. By confining the system in a slit or pore with dimensions smaller than ξ , a measurable mobility can be recovered below the bulk glass transition temperature, because structural relaxation then requires the collective rearrangement of fewer molecules than in bulk and is therefore facilitated. On the other hand, in systems confined between walls, the dynamics of the interfacial regions can be slower than that of the rest of the system, because of the interaction of the interfacial molecules with the boundaries^{24–26}. Depending on the relative strength of the effects of decreased system size and interfacial interactions, relaxation therefore become quicker or slower in confined geometries. Indeed, both decreases and increases in the glass transition temperature have been observed experimentally in confined glassy systems²⁴. To eliminate the effects stemming from this inhomogeneity of confined systems, we have used a local probe to measure the mobility of glass-forming molecules at the interface with quartz substrates in films of varying thickness. This allows us to directly probe the motion of molecules

at a definite position in the films as a function of the size of the system.

To detect the interfacial dynamics, we take advantage of the surface-specific optical second-harmonic generation technique (which probes the orientational order of liquid crystalline groups^{27,28}) by using glass-forming liquid-crystal molecules (provided by Akzo-Nobel²⁹, Arnhem, The Netherlands) (Fig. 1). As second-harmonic generation is forbidden in centro-symmetric media (in the dielectric approximation), this technique selectively probes regions where polar ordering is present; in the present case, this occurs only in the layer of liquid-crystal molecules in contact with the quartz substrate.

By spin-coating the glass-forming liquid crystal from solution, we obtain films that are not equilibrated. The polar order at the interface is then initially rather low, and hence the second-harmonic signal is small. By measuring the time-dependence of this signal as it grows towards its equilibrium value, we can follow the reorientational dynamics of the interfacial molecules as they relax towards equilibrium.

We have studied this interfacial dynamics at different temperatures below the bulk glass transition temperature in films of different thicknesses d ranging from a monolayer (30 Å) to 80 Å. The time-dependence of the second-harmonic field can be fitted to an exponential (Fig. 1): $E(t) = E_{\infty} - \Delta E \exp(-t/\tau)$ defining the relaxation time τ of the interfacial molecules. τ depends on temperature and film thickness (Fig. 2). At 294 K, only monolayers show a measurable relaxation. At 308 K and 323 K, the relaxation time is measurable over a wide enough range of thicknesses to establish that τ exponentially increases when d increases.

A model that provides a simple explanation for this exponential thickness-dependence is the configurational entropy model of glass formation of Adam and Gibbs¹⁵. This model, developed for a bulk system, yields the relaxation time τ of collectively rearranging regions (spheres of radius ξ) containing N_b molecules: $\tau = \tau_0 \exp(\alpha_b N_b)$. The term α_b represents the free-energy barrier per molecule which has to be overcome in the collective rearrangement of all molecules in a given region; it is a measure of the mobility of the molecules.

In a film of thickness d , the collectively rearranging regions are truncated in the direction normal to the film, and the number of molecules they contain is therefore smaller. Moreover the

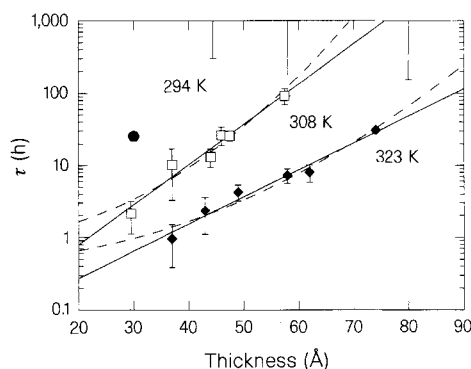


Figure 2 Thickness-dependence of the interfacial relaxation time τ at three different temperatures (294, 308 and 323 K) below the bulk glass transition temperature $T_g = 383$ K (S. J. Picken and F. G. H. van Wijk, personal communication). The measured signal became irreproducible after 3–4 days, making the measurement of relaxation times larger than 100 hours impossible. Therefore only lower limits of τ could be determined for the thickest films measured at each temperature. The lines correspond to least-square fits to two functions compatible with our data: $\tau(d, T) = \tau_0(T) \exp(d/d_0(T))$ (full lines) and $\tau(d, T) = \tau_0(T) \exp(d/d_0(T))^2$ (dashed lines).

environment of the molecules depends on their location in the film and on d , and so does their mobility. The relaxation time of molecules located at a height Z (above the substrate surface) in a film of thickness d becomes then (we neglect the weaker Z - and d -dependence of the factor τ_0 outside the exponential): $\tau(Z, d) = \tau_0 \exp(\int \alpha(z, d) n(z, d) dz)$, where the integral is taken on the collectively rearranging volume $V(Z, d)$ of the film located within a radius ξ around the considered molecules, and $n(z, d)$ is the number of molecules in $V(Z, d)$ located at the altitude z (Fig. 3).

The expression of $\alpha(z, d)$ can be obtained by an iterative self-consistent process based on the idea that the mobility of a given molecule at a height Z is essentially the average of the mobility of the other molecules located within $V(Z, d)$. As a first guess for α , we take: $\alpha(Z, d) = \alpha_b N(Z, d)/N_b + \alpha_s n_s(Z)/N_b$. The first term corresponds to a pure confinement effect: the mobility of the molecules increases when the number $N(Z, d)$ of molecules in the collectively rearranging regions decreases. The second term corresponds to the direct effect of the substrate on the molecules in contact with it (for which $Z = 0$): the mobility of these molecules might be reduced, which is translated into an extra energy barrier α_s per molecule. In principal, a similar effect exists at the free surface since this surface can hinder out-of-plane reorientations of the molecules. However the corresponding energy barrier is expected to be much smaller than the one due to the substrate, which hinders all rotational and translational degrees of freedom of the molecules. The last term corresponds to an indirect effect of the substrate on the molecules located at a distance Z from it smaller than ξ , for which $V(Z, d)$ contains a certain number $n_s(Z)$ of interfacial molecules.

This first guess turns out to be very good for $d \leq \xi$. The relaxation time $\tau(Z = 0, d)$ of the molecules located at the interface with the substrate contains then three terms associated with the three different effects just described (B.J., manuscript in preparation). The direct effect of the substrate yields a thickness-independent term. The other two terms are due to confinement and are exponential functions of the film thickness. One of these terms exists independently of the presence of a substrate and gives for small values of d a contribution to $\ln[\tau(Z = 0, d)/\tau_0]$ in $(d/\xi)^2$. This quadratic variation arises from the fact that confinement both reduces the number of molecules which have to rearrange collectively, and increases the mobility of these molecules. The other

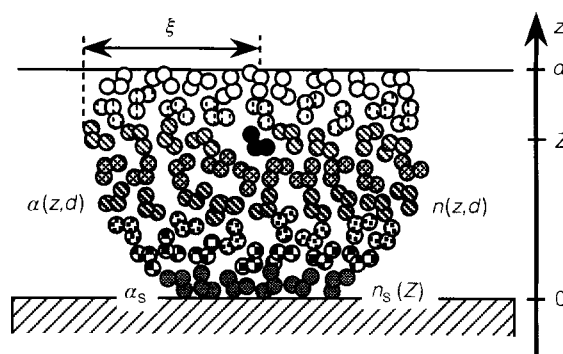


Figure 3 Schematic representation of a collectively rearranging region $V(Z, d)$ around a molecule (in black) located at height Z in a film of thickness d . In this region, there are $n(z, d)$ molecules at height z with an associated energy barrier $\alpha(z, d)$, and $n_s(Z)$ molecules in contact with the substrate with an extra energy barrier α_s .

confinement term exists only in the presence of a substrate and gives a contribution in d/ξ : it arises from the reduction of the number of molecules indirectly affected by the substrate.

So the experimentally observed exponential increase of $\tau(Z = 0, d)$ when d increases can be described by the effect of confinement on a collectively relaxing system. For film thicknesses larger than the one we have explored so far, one expects a saturation of the relaxation time. For $d > \xi$, $\tau(Z = 0, d)$ saturates at a value $\tau_0 \exp[0.45\alpha_b N_b + 1.4\alpha_s n_s(0)]$, while $\tau(Z = d, d)$ at the free surface tends towards $\tau_0 \exp[0.45\alpha_b N_b]$ and $\tau(Z = d/2, d)$ at the middle of the film tends towards the bulk relaxation time $\tau_0 \exp[\alpha_b N_b]$.

Thus, our measurements of interfacial dynamics in confined films show that the mobility of interfacial glass-forming molecules is affected by changes in the thickness of the film. The decrease in interfacial mobility with increasing film thickness (that is, increasing length scale over which collective motion takes place) results in a slowing down of the collective dynamics, eventually leading to a glass state. This provides direct experimental evidence that molecular motion in glass-forming liquids is collective: the dynamics of the interfacial molecules depends on the behaviour of the entire film. □

Received 5 December 1996; accepted 12 February 1997.

- Angell, C. A. Relaxation in liquids, polymers and plastic crystal-strong/fragile patterns and related problems. *J. Non-Cryst. Solids* **131**–**133**, 13–31 (1991).
- Jäckle, J. Models of the glass transition. *Rep. Prog. Phys.* **49**, 171–231 (1986).
- Angell, C. A. Perspective on the glass transition. *J. Phys. Chem. Solids* **49**, 863–871 (1988).
- Hansen, J. P. Clarifying the kinetic glass transition. *Phys. World* **4**(12), 32–36 (1991).
- Götze, W., Sjögren, L. Relaxation process in supercooled liquids. *Rep. Prog. Phys.* **55**, 241–376 (1992).
- Kivelson, D., Tarjus, G., Zhao, X. & Kivelson, S. A. Fitting of viscosity: distinguishing the temperature dependences predicted by various models of supercooled liquids. *Phys. Rev. E* **53**, 751–758 (1996).
- Kirkpatrick, T. R., Thirumalai, D. & Wolynes, P. G. Scaling concepts for the dynamics of viscous liquids near an ideal glassy state. *Phys. Rev. A* **40**, 1045–1054 (1989).
- Sethna, J. P., Shore, J. D. & Huang, M. Scaling theory for the glass transition. *Phys. Rev. B* **44**, 4943–4959 (1991).
- Brüning, R. & Samwer, K. Glass transition on long time scales. *Phys. Rev. B* **46**, 11318–11322 (1992).
- Jäckle, J. Thermoviscoelastic theory of freezing of stress and strain in a symmetrically cooled infinite glass plate. *J. Non-Cryst. Solids* **172**–**174**, 104–107 (1994).
- Cicerone, M. T., Blackburn, F. R. & Ediger, M. D. How do molecules move near T_g ? Molecular rotation of six probes in *o*-terphenyl across 14 decades in time. *J. Chem. Phys.* **102**, 471–479 (1995).
- Kivelson, D., Kivelson, S. A., Zhao, X., Nussinov, Z. & Targus, G. A thermodynamic theory of supercooled liquids. *Physica A* **219**, 27–38 (1995).
- Jenckel, E. Zur Temperaturabhängigkeit der Viskosität von Schmelzen. *Z. Phys. Chem.* **184**, 309–319 (1939).
- McLaughlin, E. & Ubbelohde, A. R. Structure and viscosity of melts of aromatic hydrocarbons. *Trans. Faraday Soc.* **54**, 1804–1810 (1958).
- Adam, G. & Gibbs, J. H. On the temperature dependence of cooperative relaxation properties in glass-forming liquids. *J. Chem. Phys.* **43**, 139–146 (1965).
- Fisher, E. W., Donth, E. & Steffen, W. Temperature dependence of characteristic length for glass transition. *Phys. Rev. Lett.* **68**, 2344–2346 (1992).

17. Kondo, T. & Tsumuraya, K. Isosahedral clustering in a supercooled liquid and glass. *J. Chem. Phys.* **94**, 8220–8226 (1991).
18. Mansfield, K. F. & Theodorou, D. N. Molecular dynamics simulation of a glassy polymer surface. *Macromolecules* **24**, 6283–6294 (1991).
19. Ray, P. & Binder, K. Finite-size effect in the dynamics near the glass transition. *Europhys. Lett.* **27**, 53–58 (1994).
20. Ernst, R. M., Nagel, S. R. & Grest, G. S. Search for a correlation length in a simulation of the glass transition. *Phys. Rev. B* **43**, 8070–8080 (1991).
21. Jackson, C. L. & McKenna, G. B. The glass transition of organic liquids confined to small pores. *J. Non-Cryst. Solids* **131–133**, 221–224 (1991).
22. Zhang, J., Liu, G. & Jonas, J. Effects of confinement on the glass transition temperature of molecular liquids. *J. Phys. Chem.* **96**, 3478–3480 (1992).
23. Reiter, G. Mobility of polymers in films thinner than their unperturbed size. *Europhys. Lett.* **23**, 579–584 (1993).
24. Keddie, J. L., Jones, R. A. L. & Cory, R. A. Size-dependent depression of the glass transition temperature in polymer films. *Europhys. Lett.* **27**, 59–64 (1994); Interface and surface effects on the glass-transition temperature in thin polymer films. *Faraday Discuss.* **98**, 219–230 (1994).
25. Forrest, J. A., Dalnoki-Veress, K., Stevens, J. R. & Dutcher, J. R. Effect of free surfaces on the glass transition temperature of thin polymer films. *Phys. Rev. Lett.* **77**, 2002–2005 (1996).
26. Schüller, J., Mel'nikchenko, Yu. B., Richert, R. & Fischer, E. W. Dielectric studies of the glass transition in porous media. *Phys. Rev. Lett.* **73**, 2224–2227 (1994).
27. Guyot-Sionnest, P., Hsiung, H. & Shen, Y. R. Surface polar ordering in a liquid crystal observed by optical second-harmonic generation. *Phys. Rev. Lett.* **57**, 2963–2966 (1986).
28. Jérôme, B., O'Brien, J., Ouchi, Y., Stanners, C. & Shen, Y. R. Bulk reorientation driven by orientational transit in a liquid crystal monolayer. *Phys. Rev. Lett.* **71**, 758–761 (1993).
29. de Wit, P. P., Erdhuisen, E. W. P. & Picken, S. J. patent WO 96/03476.

Acknowledgements. We thank P. P. de Wit (Akzo-Nobel) for providing the glass-forming liquid crystal; S. J. Picken, F. G. H. van Wijk (Akzo-Nobel), A. L. Demirel and B. Mulder for discussions; W. H. de Jeu for the use of the X-ray spectrometer and the spin-coater; and E. A. L. Mol and J. D. Shindler for help with the X-ray reflectivity measurements. This work is part of the research programme of the Stichting voor Fundamenteel Onderzoek der Materie (FOM), which is supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Correspondence and requests for materials should be addressed to B.J. (e-mail: jerome@amolf.nl).

Inherently unstable climate behaviour due to weak thermohaline ocean circulation

Eli Tziperman

Department of Environmental Sciences, Weizmann Institute of Science, Rehovot, 76100, Israel and Atmospheric and Oceanic Sciences Program, Geophysical Fluid Dynamics Laboratory, Princeton University, Princeton, New Jersey 08544, USA

The oceanic thermohaline circulation (THC) carries light, warm surface water polewards and dense, cold deep water equatorwards, thereby transporting a large amount of heat towards the poles and significantly affecting high latitude climate. The THC has been remarkably stable, and its variability quite low, over the Holocene period (the past 10,000 years). The much greater climate instability and high-frequency variability recorded in ice¹ and deep-sea³¹ cores throughout the preceding 150,000 years has been linked to greater THC variability^{2,3}. Here we argue, using a global coupled ocean–atmosphere–ice general circulation model with realistic geography, that there is a wide range of weak mean states of the THC that cannot be stably sustained by the climate system. When the model THC is forced into a state in the unstable range, the THC may rapidly strengthen, collapse or display strong oscillations. The existence of this unstable regime may account for the greater variability of the THC and climate before the Holocene period.

The THC is driven by the north–south gradient in water density created by atmospheric thermal forcing and weakened by the north–south salinity gradient maintained by air–sea exchange of fresh water (FW) due to subtropical evaporation and high-latitude precipitation. The present-day stable THC is thermally dominant, that is, driven by cooling and sinking of dense water masses at high latitudes. Numerous ocean-only model studies, from simple box models⁴ to three-dimensional models⁵, have shown that a weaker

THC than that of today, obtained under larger freshwater forcing (that is, a less thermally dominant THC) may become unstable. Once unstable, such a weak THC cannot be sustained with small amplitude variability such as exists today. Instead, the THC may reverse^{4,6–8} (as it may have done during the Cretaceous period⁹), undergo a significant decrease¹⁰, collapse^{10,11}, display strong oscillations^{12–14} or increase to a stable value outside the unstable range¹⁰. On the basis of such ocean-only models, it was further speculated^{6,15} that a weaker and less thermally dominant North Atlantic THC than today's, obtained under stronger FW forcing (for example, a larger FW flux into the northern North Atlantic Ocean), might be unstable.

Ocean-only models heretofore used for investigating the stability of the THC are known to be prone to various artefacts^{8,16,17} which can affect the model's stability behaviour. Here I use a coupled ocean–atmosphere–ice general circulation model (GCM), in which the air–sea interaction is more fully represented, to examine the spontaneous stability of a range of strong (thermally dominant) and weak (less thermally dominant) THC states. We consider a given climate state to be stable if, when used to initialize the coupled model, it results in a small-amplitude variability around this initial state, similar to the weak THC variability during the Holocene. On the other hand, a climate state is considered unstable if, when used to initialize the coupled model, it results in a sizeable drift of the THC away from the initial state, or in large-amplitude THC oscillations. A THC instability may thus result in an increase as

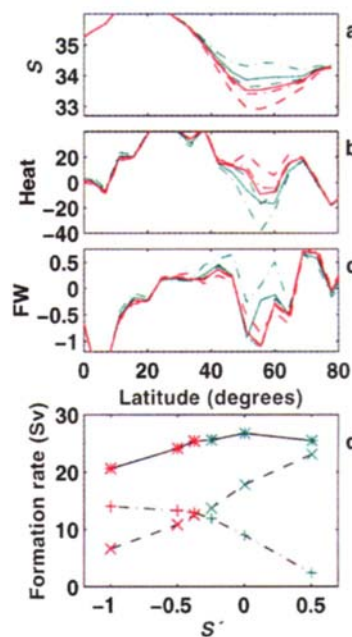


Figure 1 Diagnostics of the steady-state, ocean-only solutions used to initialize the coupled model integrations shown in Fig. 2. **a**, The zonally averaged North Atlantic sea surface salinity fields used to obtain the various steady-state, ocean-only solutions. To the observed North Atlantic salinity (solid grid line), zonally uniform perturbations of the form $\Delta S \exp(-(\theta - 55.5^\circ)^2/4.4^\circ)$ were added, as a function of latitude θ only, where ΔS takes the values $(-1, -0.5, -0.375, -0.25, 0, +0.5 \text{ p.p.t.})$. **b**, the heat flux and **c**, the freshwater flux adjustments (in W m^{-2} and m yr^{-1} , respectively), annually averaged and zonally averaged over the Northern Atlantic ocean, for the control run and for the five modified surface salinity runs. A positive FW adjustment tends to make the ocean saltier, and a positive heat adjustment heats the ocean. Line types for each experiment are as in **a**. **d**, The steady-state, ocean-only rate of formation of North Atlantic Deep Water (dashed line), Southern Ocean Deep Water (dash-dotted line) and their sum (solid line), as function of the restoring salinity perturbation, $S' = \Delta S$, in the North Atlantic.

well as a decrease, a collapse or strong oscillations of the THC¹⁰ which resemble some of the strong climate variability seen during unstable past climates¹.

To explore the stability of weak and strong (less and more thermally dominant) THC states, I initialized several coupled ocean–atmosphere model runs from various steady-state, ocean-only model solutions, following the standard initialization procedure for such coupled models¹⁸. The different ocean-only steady-state solutions were obtained by forcing (restoring) the upper level temperature and salinity of the ocean model to near an observed monthly sea surface temperature (SST) and a modified observed monthly sea surface salinity (SSS)¹² (termed ‘restoring SST and SSS’). The modifications to the SSS to which the surface model salinity is restored makes it fresher or saltier in the North Atlantic area of water mass formation (Fig. 1a), leading to a correspondingly weaker (less thermally dominant) or stronger (more thermally dominant) steady-state ocean-only North Atlantic THC (dashed line in Fig. 1d). The restoring of the model surface temperature and salinity to the specified SST and SSS fields allows various ocean-only steady states to be obtained, some of which may be unstable in the coupled model, where the surface fields are not restored and are free to evolve.

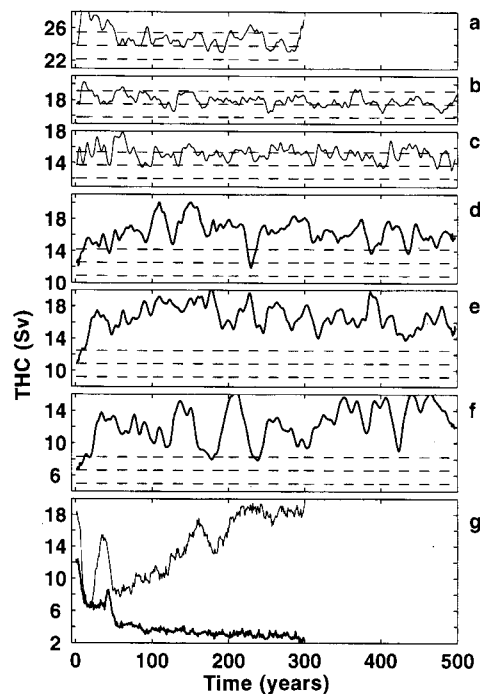


Figure 2 The North Atlantic yearly averaged THC (in Sverdrup) plotted against time (years) from eight coupled model runs. Dashed lines denote the initial state and plus/minus two standard deviation (SD) lines, where the SD is defined from the control run (b). Stable initial states are defined as those whose THC remains within two SD during the coupled model run, although the results are not sensitive to this particular choice of two SD. The different panels show **a**, the coupled model run initialized with the stable state [$\Delta S = +0.5$]; **b**, the stable control run^{19,20}; **c**, run initialized with the stable state [$\Delta S = -0.25$]; **d**, run initialized with the unstable state [$\Delta S = -0.375$]; **e**, run initialized with the unstable state [$\Delta S = -0.5$]; **f**, run initialized with the unstable state [$\Delta S = -1$]; **g**, thick line: the North Atlantic THC for a coupled model run starting from year 10 of the run shown in **e**, initialized with the unstable initial state [$\Delta S = -0.5$]. In this run, freshwater flux is added to the North Atlantic sinking area for the first 10 years. The THC collapses and does not recover for 300 years. This contrasts with the corresponding experiment of ref. 23, started from the stable control run and shown by the thin line.

The low-resolution primitive-equation coupled ocean–atmosphere–ice GCM used here, with its realistic geometry, has been used for many studies of greenhouse warming¹⁹, climate variability²⁰ and multiple climate equilibria²¹. When initialized with ocean-only and atmosphere-only steady states obtained by running the two models separately, the coupled model rapidly drifts from its initial conditions. The air–sea fluxes of heat and fresh water are therefore supplemented at every time step by flux adjustments (FA) which depend on geographical location and month, but have no inter-annual variations and are obtained by averaging the restoring terms from the ocean-only initialization run¹⁸. FA is used in the present experiments as a parameter that allows the different initial THC states to be obtained (Fig. 1b, c). This is somewhat similar to changing the FW forcing in an ocean-only model^{12,14,15}.

In the absence of an external perturbation, if the initial THC is stable, we expect the coupled-model initialization procedure to prevent a drift of the THC in the coupled model from the initial ocean-only steady states, as it did in the stable control run^{19,20} (although see ref. 22 for a caveat). The strength of the North Atlantic THC as a function of time for the five coupled-model runs initialized using modified surface salinities, and for the control run, is shown in Fig. 2a–f. I denote each initial state of the coupled model by the amplitude ΔS of the restoring salinity perturbation used to obtain it in the ocean-only model runs (Fig. 1a). When ocean-only models in previous studies were initialized by restoring their SST and SSS and were then switched to some simple atmospheric representation, these models invariably showed that weak THC initial states were unstable, leading to a switch to a different THC state. The qualitative behaviour of the realistic-geometry coupled ocean–atmosphere GCM used here is remarkably consistent with that of these simpler ocean-only models. In particular, the more thermally dominant initial states with a strong initial THC (runs [$\Delta S = -0.25$], control [$\Delta S = 0$] and [$\Delta S = +0.5$], Fig. 2a–c) are stable. The coupled model runs starting at these initial conditions remain mostly within two standard deviations from the initial state, where the standard deviation is defined from the control run's THC variability which roughly represents a small, Holocene-like, stable variability. On the other hand, the less thermally dominant and thus weaker initial THC states (runs [$\Delta S = -1$], [$\Delta S = -0.5$], [$\Delta S = -0.375$], Fig. 2d–f) are spontaneously unstable within the coupled model, as the THC drift from the initial conditions in these runs clearly exceeds two standard deviations. I call this a spontaneous instability, because no external perturbation is applied to the coupled model once the coupled model integration begins.

The threshold between stable and spontaneously unstable initial states in the coupled model lies between 14 Sv (stable state [$\Delta S = -0.25$], Fig. 2c) and 13 Sv (unstable state [$\Delta S = -0.375$], Fig. 2d). The steady-state ocean-only solution used as the stable initial condition in the control run [$\Delta S = 0$] is different from the initial state [$\Delta S = -0.375$] that is unstable within the coupled model, yet not dissimilar (Fig. 3a–d), indicating that an ocean climate not very far from the present-day climate may be spontaneously unstable within the coupled model. Note that none of the runs displays a stable small-amplitude THC variability, like that of the present day, within the unstable THC regime of below 14 Sv. When forced by a doubling of carbon dioxide¹⁹ and by freshwater input into the North Atlantic ocean²³, the coupled model THC recovers within 150–200 years. Its inability to recover to its original weak initial states within the fairly long integration of 500 years seen in Fig. 2 indicates that the weak THC initial states may indeed be considered to be in an unstable THC regime.

The instability mechanism leading to the drift away from the unstable initial states in Fig. 2d–f is composed of two stages. First, an advective linear salinity feedback^{6,7,15} leads to an increase of salinity and density in the water mass formation area and thus an increase in the formation rate of DW southeast of Greenland. The

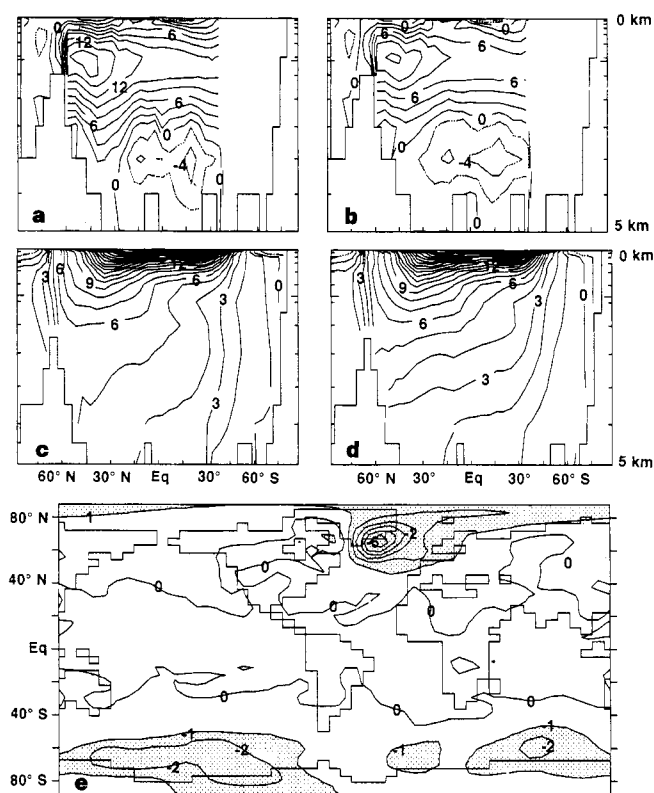


Figure 3 a–d, Upper panel, the steady-state, ocean-only North Atlantic meridional circulation (that is, the meridional stream function); middle panel, the zonally averaged temperature. **a, c**, Stable initial state for the control run; **b, d**, unstable initial state for run $[\Delta S = -0.375]$. The smaller rate of deep-water formation in the North Atlantic in the unstable solution allows an increased influx of Southern Ocean water at depth, and thus the difference seen in the deep-water properties and in Deep Water formation rate in the Southern Ocean (Fig. 1d). **e**, The difference between the atmospheric surface temperature averaged over years 201–300 for the collapsed THC run of Fig. 2g, and the similarly averaged atmospheric surface temperature for the control run.

runs with the stronger initial THC are stable with respect to this linear instability stage, because the strong THC rapidly advects away any developing salinity perturbations, and because such perturbations are harder to develop because of the reduced meridional salinity gradient in the stable states^{6,7,15}. At a second stage, the increased sinking southeast of Greenland modifies the properties of the deep water inside the neighbouring Labrador Sea, activating DW formation and self-sustained convective mixing²⁴ there as well, and leading to a further increase in the THC strength. Such an increase is not seen in simpler two-dimensional models^{4,7,15} which, once unstable, often lead to a collapsed or reversed THC. These models, being two-dimensional, cannot support an inactive DW formation site at the same latitude as the active formation sites.

Past the initial instability, the runs beginning from the unstable initial states are characterized by strong North Atlantic THC oscillations (Fig. 2d–f). This is the first time that such large-amplitude oscillations have been seen in a coupled model, following numerous ocean-only model studies that observed such large-amplitude THC variability⁵, and coupled model runs that displayed a smaller variability^{20,25}. These large-amplitude oscillations (which, especially in Fig. 2f, seem to result from the inherent instability of the weak THC states) do not involve a complete switch between 'on' and 'off' states of the North Atlantic THC, yet may be able to account for the climate oscillations in the interglacial Eemian period^{1,26}.

In a final experiment (Fig. 2g), I added a freshwater flux of 1 Sv to the northern North Atlantic over 10 years, and found that the spontaneously unstable initial states were also more prone to a forced THC collapse than stronger THC runs such as the control run²³.

Various palaeo proxy data have shown a link between the North Atlantic THC and the Pacific sector of the Southern Ocean². The present ocean-only steady-state solutions indeed show that as DW formation rates in the Pacific sector of the Southern Ocean and in the North Atlantic Ocean vary between the different initial states by 10–15 Sv at each DW source, their sum remains nearly constant (Fig. 1d). Thus this sum, the total DW formation, is determined through a global teleconnection (Fig. 3). Similarly, the forced instability experiment (Fig. 2g) displays a cooling of the air temperature both over Greenland and in the Pacific sector of the Southern Ocean by about 5 and 2 °C respectively (Fig. 3e; see also ref. 23).

The unknown effects of the flux adjustment procedure introduce an uncertainty as to the applicability of the more quantitative results (the precise location of the instability threshold) to the actual climate system^{27,28}. I also note that the phenomenon of a coupling shock²², at the initialization of coupled model integrations, may lead to an initial model drift which may affect the differentiation of stable and unstable states. However, some previous coupled model results²⁹, and in particular the remarkable qualitative agreement with the numerous ocean-only studies of THC stability lend additional confidence to the results here. In fact, I feel that a flux-adjusted coupled model with an active atmospheric component may be expected to be more accurate than ocean-only models with fixed freshwater flux and atmospheric temperature, with a highly simplified heat flux parameterization and with no weather noise.

In summary, this study provides confirmation using a fully coupled ocean–atmosphere–ice general circulation model with realistic geometry that a weak, less thermally dominant THC is unstable. The results limit the range of stable equilibria of the THC in this model to above about 14 Sv and below 4 Sv (our collapsed THC run, Fig. 2g). In this model I did not find intermediate values of the THC that are stable, with small variability like that of the present day. This may suggest that the North Atlantic water mass formation and thus THC may have operated in a weak and shallow³ (Fig. 3b) unstable mode with large variability before the Holocene and during previous interglacial periods¹. Later, the water mass formation may have switched to a stable mode with small variability, and with stronger and deeper North Atlantic THC, which has lasted throughout the Holocene.

To set this in perspective, I note that the weakening required to make the model THC spontaneously unstable is still significantly larger than the present-day natural variability, which is estimated to be about 5–10% for the coupled model used here²⁰. Although the location of the stable THC range found here differs from those found in simpler models^{15,21} and is clearly model dependent, I expect the qualitative dependence of the model stability and variability behaviour on the mean THC strength to carry over to the real world. The implications for greenhouse scenarios that predict a weakening of the THC are clearly an important yet open issue. Finally, improved knowledge of past THC magnitude³⁰, and a more precise determination of the unstable THC regime and its distance from present-day climate, are clearly needed, using better coupled ocean–atmosphere models as they become available. □

Received 10 October 1996; accepted 7 February 1997.

1. GRIP Members Climate instability during the last interglacial period recorded in the GRIP ice core. *Nature* **364**, 203–207 (1993).
2. Broecker, W. S., Petet, D. M. & Rind, D. Does the ocean–atmosphere system have more than one stable mode of operation? *Nature* **315**, 21–25 (1985).
3. Boyle, E. A. & Keigwin, L. North Atlantic thermohaline circulation during the past 20,000 years linked to high-latitude surface temperature. *Nature* **330**, 35–40 (1987).
4. Stommel, H. M. Thermohaline convection with two stable regimes of flow. *Tellus* **XIII**, 224–230 (1961).

5. McWilliams, J. C. Modelling the oceanic general circulation. *Annu. Rev. Fluid. Mech.* **28**, 215–248 (1996).
6. Walin, G. The thermohaline circulation and the control of ice ages. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **50**, 323–332 (1985).
7. Marotzke, J., Welander, P. & Willebrand, J. Instability and multiple steady states in a meridional-plane model of the thermohaline circulation. *Tellus A* **40**, 162–172 (1988).
8. Mikolajewicz, U. & Maier-Reimer, E. Mixed boundary conditions in ocean general circulation models and their influence on the stability of the model's conveyor belt. *J. Geophys. Res.* **99**, 22633–22644 (1994).
9. Brass, G. W., Southam, J. R. & Peterson, W. H. Warm saline bottom water in the ancient ocean. *Nature* **296**, 620–623 (1982).
10. Toggweiler, J. R. *et al.* Reply. *J. Phys. Oceanogr.* **26**, 1106–1110 (1996).
11. Bryan, F. High-latitude salinity effects and interhemispheric thermohaline circulations. *Nature* **323**, 301–304 (1986).
12. Weaver, A. J., Sarachick, E. S. & Marotzke, J. Freshwater flux forcing of decadal and interdecadal oceanic variability. *Nature* **353**, 836–838 (1991).
13. Chen, F. & Ghil, M. Interdecadal variability of the thermohaline circulation and high-latitude surface fluxes. *J. Phys. Oceanogr.* **25**, 2547–2568 (1995).
14. Rahmstorf, S. Bifurcations of the Atlantic thermohaline circulation in response to changes in the hydrological cycle. *Nature* **378**, 145–149 (1995).
15. Tziperman, E., Toggweiler, R., Feliks, Y. & Bryan, K. Instability of the thermohaline circulation with respect to mixed boundary conditions: Is it really a problem for realistic models? *J. Phys. Oceanogr.* **24**, 217–232 (1994).
16. Zhang, S., Greatbatch, R. J. & Lin, C. A. A reexamination of the polar halocline catastrophe and implications for coupled ocean–atmosphere modeling. *J. Phys. Oceanogr.* **23**, 287–299 (1993).
17. Rahmstorf, S. & Willebrand, J. The role of temperature feedback in stabilizing the thermohaline circulation. *J. Phys. Oceanogr.* **25**, 787–805 (1995).
18. Manabe, S., Spelman, M. J. & Bryan, K. Transient response of a coupled ocean–atmosphere model to gradual changes of atmospheric CO₂. Part I: Annual mean response. *J. Clim.* **4**, 785–818 (1991).
19. Manabe, S. & Stouffer, R. J. Century-scale effects of increased atmospheric CO₂ on the ocean–atmosphere system. *Nature* **364**, 215–218 (1993).
20. Delworth, T., Manabe, S. & Stouffer, R. J. Interdecadal variations of the thermohaline circulation in a coupled ocean–atmosphere model. *J. Clim.* **6**, 1993–2011 (1993).
21. Manabe, S. & Stouffer, R. J. Two stable equilibria of a coupled ocean–atmosphere model. *J. Clim.* **1**, 841–866 (1988).
22. Rahmstorf, S. Climate drift in an ocean model coupled to a simple, perfectly matched atmosphere. *Clim. Dynam.* **11**, 447–458 (1996).
23. Manabe, S. & Stouffer, R. J. Simulation of abrupt climate change induced by freshwater input to the North Atlantic ocean. *Nature* **378**, 165–167 (1995).
24. Lenderink, G. & Haarsma, R. J. Variability and multiple equilibria of the thermohaline circulation associated with deep water formation. *J. Phys. Oceanogr.* **24**, 1480–1493 (1994).
25. Griffies, S. & Tziperman, E. A linear thermohaline oscillator driven by stochastic atmospheric forcing. *J. Clim.* **8**, 2440–2453 (1995).
26. Weaver, A. J. & Hughes, T. M. C. Rapid interglacial climate fluctuations driven by North Atlantic ocean circulation. *Nature* **367**, 447–450 (1994).
27. Neelin, J. D. & Dijkstra, H. A. Ocean–atmosphere interaction and the tropical climatology. Part I: the dangers of flux adjustment. *J. Clim.* **8**, 1325–1342 (1995).
28. Marotzke, J. & Stone, P. H. Atmospheric transports, the thermohaline circulation, and flux adjustments in a simple coupled model. *J. Phys. Oceanogr.* **25**, 1350–1364 (1995).
29. Schiller, A., Mikolajewicz, U. & Voss, R. The stability of the thermohaline circulation in a coupled ocean–atmosphere general circulation model. *Clim. Dynam.* (in the press).
30. Yu, E.-F., Francois, R. & Bacon, M. P. Similar rates of modern and last-glacial ocean thermohaline circulation inferred from radio-chemical data. *Nature* **379**, 689–694 (1996).
31. McManus, J. F. *et al.* High resolution climate records from the North Atlantic during the last interglacial. *Nature* **371**, 326–329 (1995).

Acknowledgements. I thank S. Manabe and R. Stouffer for allowing the use of their coupled model, and S. Griffies and R. Stouffer without whose help the completion of this work would not have been possible. This work is partially funded by the Israeli Academy of Sciences.

Correspondence should be addressed to E.T. at the Weizmann Institute of Science (e-mail: eli@beach.weizmann.ac.il).

Silent fault slip following an interplate thrust earthquake at the Japan Trench

Kosuke Heki*, Shin'ichi Miyazaki† & Hiromichi Tsuji‡

* National Astronomical Observatory, 2–12 Hoshigaoka, Mizusawa-city, Iwate 023, Japan

† Geographical Survey Institute, 1 Kitasato, Tsukuba-city, Ibaraki 305, Japan

‡ Ministry of Construction, 2–1–3 Kasumigaseki, Chiyoda-ku, Tokyo 100, Japan

Recent global space geodetic measurements have revealed that the velocities of tectonic plates over timescales as short as a decade¹ are consistent with models of velocities averaged over the past few million years. The slip inferred from interplate thrust earthquakes at deep sea trenches and number of earthquakes, however, often falls short of that predicted from these observed plate convergence rates^{2,3}. Here we report transient crustal movements

recorded by a permanent Global Positioning System (GPS) network in northeastern Japan following a typical interplate earthquake that occurred in December 1994 at the Japan Trench. Cumulative fault slip was estimated from the postseismic displacements at the GPS points over the first year after the event, and the inferred amount of seismic moment released by the afterslip was comparable to that released in the high-speed rupture. Such seismically ‘invisible’ slip may therefore account for the shortage of seismic slip relative to that required by time-averaged plate velocities.

Slips associated with modern and historical interplate earthquakes at the Japan Trench (off the Sanriku coast, northeastern Japan) account for only one-fifth of the time-averaged rate of the Pacific plate subduction². Kawasaki *et al.*³ recently discovered there an ‘ultra-slow earthquake’ with source time of ~1 day in extensometer records and hypothesized that the deficiency is taken up by such slow processes which are difficult to detect by conventional seismometric observations. The Sanriku–Haruka–Oki earthquake (moment magnitude $M_w = 7.6$) occurred there on 28 December 1994 as a typical interplate thrust event (Fig. 1). The hypocentral depth is very shallow and the aftershock area of about 150×70 km (east–west $\times$ north–south) dips to the west along the boundary between the subducting slab and the overriding plate. The focal mechanism solutions indicate that major slip occurred in the central part of the aftershock area^{4,5}, which is consistent with the slip distribution (Fig. 1) inferred by joint inversion⁶ of tsunami waveforms and coseismic displacements observed by the Japanese nationwide permanent GPS network.

This GPS network was established in October 1994 by the Geographical Survey Institute (GSI) of Japan with about 200 dual-band P-code receivers that cover most parts of Japan⁷. We analysed the data of 16 stations located in northeastern Honshu and southern Hokkaido, October 1994 to December 1995; we used release 9.28 of the GAMIT software⁸ with International GPS Service for Geodynamics (IGS) precise ephemerides and Earth orientation parameters from the International Earth Rotation Service (IERS) Bulletin B. Tropospheric delays are estimated at each station in every 3-hour period. Daily horizontal site positions with respect to the Tsukuba IGS station ~500 km south of the epicentre (Fig. 1), were used to study postseismic crustal movements. Figure 2 shows the time series of three sites, Mutsu, Aomori and Kuji; the locations of these sites are shown in Fig. 1. Their coordinates change little before the earthquake (time $t = -0.2$ to 0.0 yr), then coseismic displacements are clearly seen as discontinuities at $t = 0$ (maximum displacement of 9.2 cm at Kuji). Postseismic crustal movements are characterized by steep onsets and gradual decay over $t = 0.0$ to 1.0 yr (maximum displacement of 5.8 cm after 1 year at Kuji).

Thatcher *et al.*⁹ derived a viscosity of 1×10^{19} Pa s for the uppermost asthenosphere in this region, which implies that viscous relaxation governs processes only with timescales of a few tens of years or longer. We assume here that the observed postseismic crustal movements are generated by a slow afterslip somewhere on the fault plane that ruptured in the mainshock, and we rule out viscous effects. Direct field observation of the fault afterslip¹⁰ and theoretical studies based on rate- and state-dependent friction laws¹¹ suggest that afterslip obeys logarithmic decay approximated by $\alpha \ln(\beta t + 1)$; this equation is a simplified form of equation (5) in ref. 11, where t is the time after an earthquake, and α and β are parameters defining overall amplitude and temporal decay, respectively. (β reflects, for example, coseismic rupture time, degree of velocity-strengthening and proportion of the stable/unstable fields; ref. 12.) We estimated α and β from GPS point horizontal displacement time series using iterative nonlinear curve-fitting techniques. We estimated α as arc parameters for individual components/sites (actually we estimated one-year cumulative displacements which are proportional to α) while β was estimated as the common global parameter because surface crustal movements would have the same

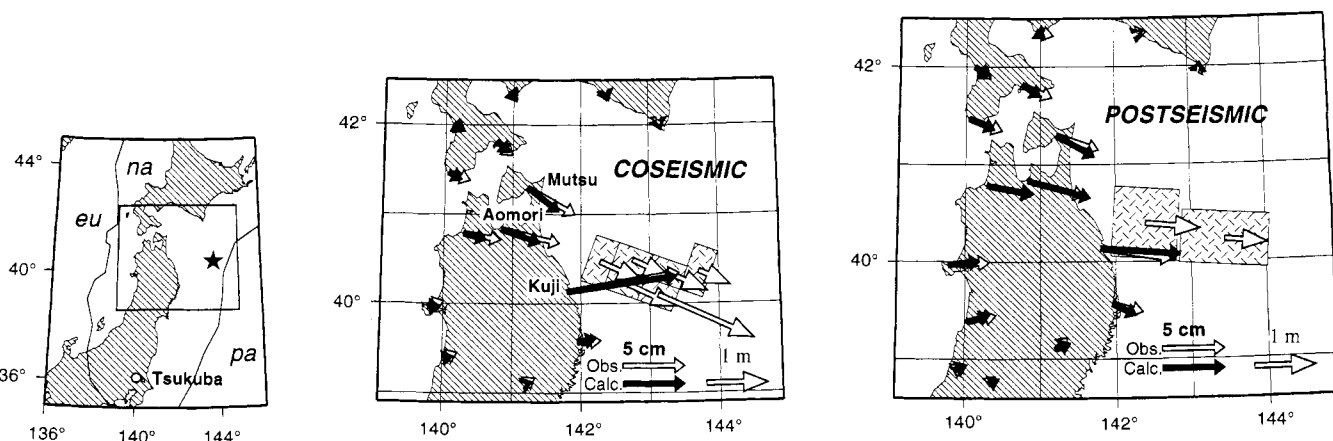


Figure 1 Left, panel showing the epicentre of the mainshock of the 28 December 1994 Sanriku-Haruka-Oki earthquake (star), plate boundaries (na, North America; eu, Eurasia; pa, Pacific), the Tsukuba GPS station (open circle), and the studied area (rectangle). The middle and right panels show respectively the co- and post-seismic crustal movements. White arrows for GPS points denote the observed horizontal movements; black arrows show the calculated movements. The 'observed' co- and post-seismic displacements were derived as discontinuities at $t = 0$ in the coordinate time series, and by fitting logarithmic decay curves to the coordinate time series over the one-year period, respectively (Fig. 2).

The 'calculated' co- and post-seismic displacements were obtained by assuming the geometry/slip of the faults shown in the figure as cross-hatched rectangles/arrows, and an elastic half-space. Coseismic fault parameters are taken from ref. 6 while postseismic slips were estimated in this study. The afterslip for the eastern segment was 69.4 ± 18.9 cm towards N97.2E (strike, 185°; dip 8°; 60×100 km; fault centre at $40^\circ 23' \text{ N}$, $143^\circ 45' \text{ E}$; depth, 17.0 km). The afterslip for the western segment was 88.5 ± 8.0 cm towards N97.2E (strike, 185°; dip, 35°; 80×90 km; fault centre at $40^\circ 39' \text{ E}$, $142^\circ 40' \text{ E}$; depth, 49.8 km).

temporal decay property as the afterslip at depth. Using 3,840 measurements from 11 stations, β was estimated as $6.5 \pm 0.6 \text{ yr}^{-1}$, that is, about half of the motion of the first year is achieved in the first 100 days. This is somewhat longer than reported in the strike-slip cases¹¹ where half of the first-year movement is often achieved in a few tens of days. We obtained a post-fit residual of 4.4 mm, a value not so different from those we usually obtain in fitting lines to the secular movements of GPS stations in Japan. Displacements of an additional five stations, farthest from the epicentre with displacements of a few millimetres, were obtained by simple linear regression.

Horizontal displacement vectors over the one-year postseismic period of the 16 GPS sites were used to derive the amount of fault slip using the dislocation theory in an elastic half-space formulated by Okada¹². Vertical components were not used because the largest expected displacement (one-year cumulative uplift of ~ 1 cm at Kuji) is as small as their day-to-day repeatability. It was almost impossible to reproduce the displacement directions when using only the afterslip in the up dip and/or down dip extensions of the aftershock area, so we assumed that most afterslip occurred in the same part of the rupture plane as the mainshock (although we cannot rule out minor afterslips in its down dip and/or up dip extensions). We used the aftershock distribution¹³ to delimit the size and the location of the faulted slip surface and divided it into two segments, the shallowly dipping eastern part and the moderately dipping western part (Fig. 1). Because GPS points exist only on one side of the fault (the land area), interparameter correlation makes it difficult to resolve detailed slip distribution by dividing it into small segments (if we split the eastern segment into two, their correlation becomes $>95\%$ and the formal errors exceed 1 metre). Slip direction could be fixed to an *a priori* value if we knew on which plate northeastern Japan resides. Because this is not the case at the moment¹⁴, we estimated the slip direction common for the two segments, together with the individual slips.

The weighted root-mean-squares of the displacement after estimating these three parameters (the slip amounts of the eastern and western fault segments, and their direction) was 4.3 mm and the estimated slip direction was N97.2E (Fig. 1) with 1σ formal error of 2.6° . This deviates by $\sim 15^\circ$ ($\sim 10^\circ$) anticlockwise from the direction

that the NUVEL1 model¹⁵ predicts for the North American (Eurasian)–Pacific plate convergence there. The estimated fault slips were 88.5 ± 8.0 cm and 69.4 ± 18.9 cm for the western and the eastern segments, respectively. Assuming⁶ a shear modulus of 40 GPa, they correspond to the seismic moment of 4.2×10^{20} N m, larger than the total moments released by the aftershocks¹³ by an order of magnitude, indicating that the slip was largely aseismic. Fit is relatively poor (that is, observed eastward velocities are too large) for the stations along the Japan Sea coast. The amount and the sense of these misfits seem consistent with the interseismic east–west shortening in this region¹⁶ because the easternmost sites of the network are situated at similar distances from the trench axis to Tsukuba, the fixed reference. If we eliminate interseismic east–west shortening of $3 \times 10^{-8} \text{ yr}^{-1}$ (ref. 16) fixing Kuji, the weighted root-mean-squares of the residuals decreases to 3.5 mm. This, however, does not change the estimated seismic moment more than a few per cent.

In October 1994, the Hokkaido–Toho–Oki earthquake⁷, an 'intraplate' earthquake, occurred within the subducting slab ~ 400 km northeast of the Sanriku–Haruka–Oki earthquake. Although it brought larger coseismic movements of GPS points in Hokkaido⁷, our preliminary analyses show that the postseismic movements are insignificant. This suggests that the distinct afterslip was peculiar to 'interplate' earthquakes, and attributable to the nature of the fault surface such as the existence of unconsolidated sediments. Pacheco *et al.*² supposed that convergent plate interfaces are made up of three parts, patches of unstable (velocity-weakening) field (that is, asperities), stable (velocity-strengthening) field and compliant (conditionally stable) field, and earthquakes nucleate only in the first field. The unstable field may be dominant in 'strongly coupled' subduction zones such as Chile, but exist only as small patches in 'weakly coupled' zones such as Sanriku where seismic coupling coefficients are small. In the present case, the major part of the after slip seems to have occurred in the same part as the coseismic rupture, in contrast with the two cases in Chile where afterslips were suggested in the up dip¹⁷ or down dip¹⁸ extensions. This case might be understood as a sequence with the instantaneous coseismic slip 'mainly' in the unstable field (that is, it might have propagated into the stable field), and the postseismic slow fault

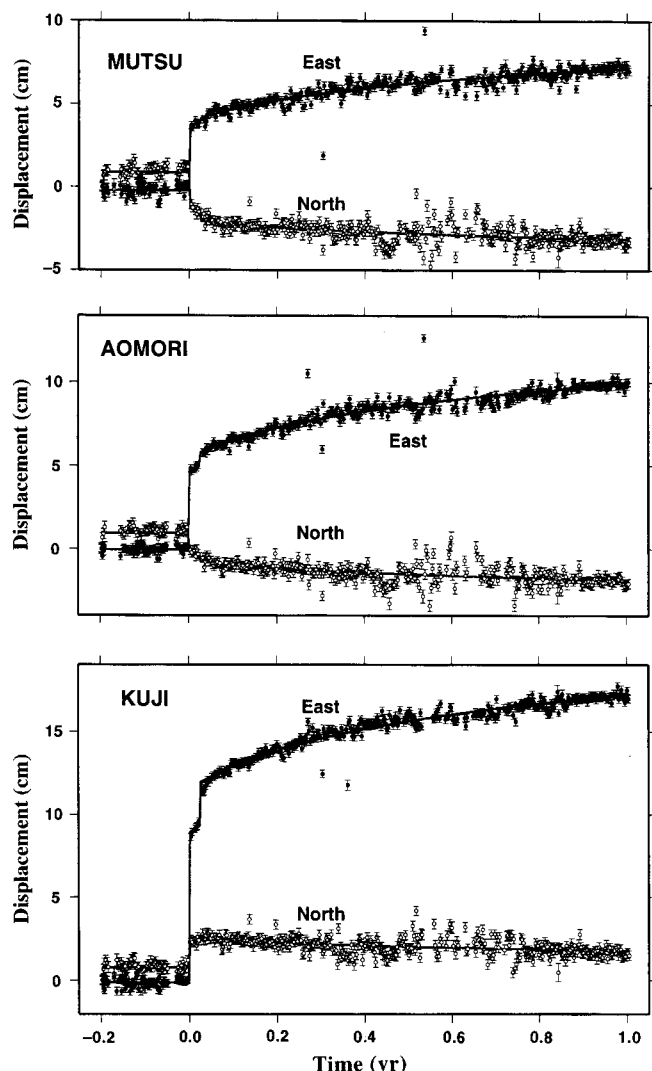


Figure 2 Horizontal coordinate time series before and after the 1994 Sanriku-Haruka-Oki earthquake observed at the three GPS points, Mutsu, Aomori and Kuji (Fig. 1). Horizontal axes show the time in years after the earthquake, and coordinates in the vertical axes are taken relative to arbitrary values. Open and filled circles denote north and east components, respectively. Thick black lines are the model curves (stationary for $t < 0$, logarithmic decay for $t > 0$, discontinuity at $t = 0.0$). The largest aftershock (surface-wave magnitude $M_s = 6.9$) occurred ~ 9.5 days after the mainshock ($t \approx 0.03$) and these stations underwent horizontal coseismic displacements of about 2.6 cm, 0.9 cm and 0.7 cm, respectively, which were removed from the curve-fitting processes by estimating additional discontinuities. No sizeable coseismic displacements were found for other smaller aftershocks (Japan Meteorological Agency counted 22 aftershocks with $M > 5$; ref. 13).

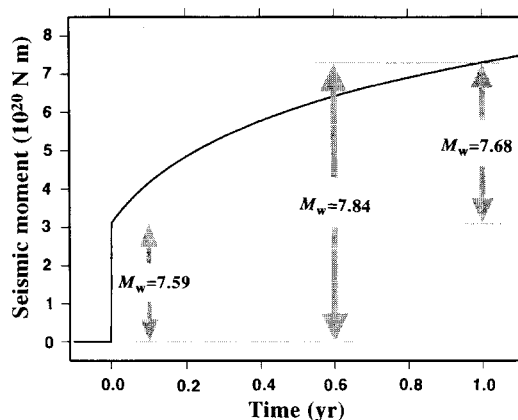


Figure 3 Coseismic and postseismic moment release curve inferred for the 1994 Sanriku-Haruka-Oki earthquake sequence. Moment magnitudes (M_w) equivalent to the coseismically released seismic moments, and the total moment after one year, are indicated.

movement 'mainly' in the compliant/stable fields with complementary slip distribution that equalizes the total slip throughout the fault plane. Although the present GPS network does not have enough resolution to derive detailed postseismic slip distribution, we can notice some differences between the coseismic and postseismic movements in Fig. 1; for example, the coseismic displacement of Kuji is more than twice as large as Aomori and deflected anticlockwise by $\sim 30^\circ$ from it, while postseismic displacement vectors of these points are more alike. This suggests that the afterslip distribution is relatively even throughout the fault surface while the coseismic slip concentrates more or less in a small central part corresponding to the asperity.

Postseismic moment release of 4.2×10^{20} N m corresponds to $M_w = 7.7$, larger than the coseismic release of 3.1×10^{20} N m ($M_w = 7.6$) suggested by the slip distribution of ref. 6. The whole sequence corresponds to an earthquake of $M_w = 7.8$ with a source time of one year (Fig. 3), much longer than ~ 1 day found in the same region³, and ~ 1 week found in California¹⁹. Together with the 1992 ultra-slow earthquake³, and the 1896 Tsunami earthquake²⁰, this case shows that it is not unusual in this region for a slow fault slip to follow a brittle fracture. Small seismic coupling coefficients found worldwide in deep sea trenches² suggest that interplate thrust earthquakes have an aspect invisible to conventional seismometric observations. Geodetic measurements of crustal deformation (especially those by GPS with sufficient

temporal and spatial coverage/density) will be important in understanding interplate thrust earthquakes over a wide range of time-scales and, above all, in allowing proper seismic hazard assessments based on seismic moment budget. $\square$

Received 11 November 1996; accepted 14 February 1997.

1. Heki, K. Horizontal and vertical crustal movements from three-dimensional very long baseline interferometry kinematic reference frame: implication for the reversal timescale revision. *J. Geophys. Res.* **101**, 3187–3198 (1996).
2. Pacheco, J. F., Sykes, L. R. & Scholz, C. H. Nature of seismic coupling along simple plate boundaries of the subduction type. *J. Geophys. Res.* **98**, 14133–14159 (1993).
3. Kawasaki, I. et al. The 1992 Sanriku-Oki, Japan, ultra-slow earthquake. *J. Phys. Earth* **43**, 105–116 (1995).
4. Sato, T., Imanishi, K. & Kosuga, M. Three-stage rupture process of the 28 December 1994 Sanriku-oki earthquake. *Geophys. Res. Lett.* **23**, 33–36 (1996).
5. Nishimura, T., Nakahara, H., Sato, H. & Ohtake, M. Source process of the 1994 far east off Sanriku earthquake, Japan, as inferred from a broad-band seismogram. *Tohoku Geophys. J.* **34**, 121–134 (1996).
6. Tanioka, Y., Ruff, L. & Satake, K. The Sanriku-oki, Japan, earthquake of December 28, 1994 ($M_w = 7.7$): rupture of a different asperity from a previous earthquake. *Geophys. Res. Lett.* **23**, 1465–1468 (1996).
7. Tsuji, H., Hatanaka, Y., Sagiya, T. & Hashimoto, M. Coseismic crustal deformation from the 1994 Hokkaido-Toho-Oki earthquake monitored by a nationwide continuous GPS array in Japan. *Geophys. Res. Lett.* **22**, 1669–1672 (1995).
8. Dong, D. & Bock, Y. Global Positioning System network analysis with phase ambiguity resolution applied to crustal deformation studies in California. *J. Geophys. Res.* **94**, 3949–3966 (1989).
9. Thatcher, W., Matsuda, T., Kato, T. & Rundle, J. B. Lithospheric loading by the 1896 Riku-u earthquake, Northern Japan: implications for plate flexure and asthenospheric rheology. *J. Geophys. Res.* **85**, 6429–6435 (1980).
10. Bucknam, R. C., Plafker, G. & Sharp, R. V. Fault movement (afterslip) following the Guatemala earthquake of February 4, 1976. *Geology* **6**, 170–173 (1978).
11. Marone, C. J., Scholtz, C. H. & Bilham, R. On the mechanics of earthquake afterslip. *J. Geophys. Res.* **96**, 8441–8452 (1991).
12. Okada, Y. Internal deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am.*

82, 1018–1040 (1992).

13. Sendai District Meteorological Observatory, Japan Meteorological Agency *The 1994 Far Off Sanriku Earthquake* (December 28, M7.5). 75–83 (Rep. 54, Coordinating Committee for Earthq. Pred., Tsukuba-city, Japan, 1995).
14. Seno, T., Sakurai, T. & Stein, S. Can the Okhotsk plate be discriminated from the North American plate? *J. Geophys. Res.* **101**, 11305–11315 (1996).
15. DeMets, C., Gordon, R. G., Argus, D. F. & Stein, S. Current plate motions. *Geophys. J. Int.* **101**, 425–478 (1990).
16. Sheng-Tu, B. & Holts, W. E. Interseismic horizontal deformation in northern Honshu and its relationship with the subduction of the Pacific plate in the Japan trench. *Geophys. Res. Lett.* **22**, 3103–3106 (1996).
17. Barrientos, S. E. Dual seismogenic behavior: the 1985 Central Chile earthquake. *Geophys. Res. Lett.* **22**, 3541–3544 (1995).
18. Barrientos, S. E., Plafker, G. & Lorca, E. Postseismic coastal uplift in southern Chile. *Geophys. Res. Lett.* **19**, 701–704 (1992).
19. Linde, A. T., Gladwin, M. T., Johnston, M. J. S., Gwyther, R. L. & Bilham, R. G. A slow earthquake sequence on the San Andreas fault. *Nature* **383**, 65–68 (1996).
20. Tanioka, Y. & Satake, K. Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys. Res. Lett.* **23**, 1549–1552 (1996).

Acknowledgements. We thank N. Kato for comments; A. Linde and P. Segall for critical reviews; and GSI staff for maintaining the GRAPES GPS network.

Correspondence should be addressed to K.H. (e-mail: heki@miz.nao.ac.jp)

Compression of visual space before saccades

John Ross*, M. Concetta Morrone† & David C. Burr†‡

* Department of Psychology, Vision Laboratory, University of Western Australia, Nedlands, Western Australia 6907, Australia

† Istituto di Neurofisiologia del CNR, Pisa 56127, Italy

‡ Department of Psychology, Università di Roma, 'La Sapienza', Via dei Marsi 78, Rome 00185, Italy

Saccadic eye movements, in which the eye moves rapidly between two resting positions, shift the position of our retinal images. If our perception of the world is to remain stable, the visual directions associated with retinal sites, and others they report to, must be updated to compensate for changes in the point of gaze. It has long been suspected that this compensation is achieved by a uniform shift of coordinates driven by an extra-retinal position signal^{1–3}, although some consider this to be unnecessary^{4–6}. Considerable effort has been devoted to a search for such a signal and to measuring its time course and accuracy. Here, by using multiple as well as single targets under normal viewing conditions, we show that changes in apparent visual

direction anticipate saccades and are not of the same size, or even in the same direction, for all parts of the visual field. We also show that there is a compression of visual space sufficient to reduce the spacing and even the apparent number of pattern elements. The results are in part consistent with electrophysiological findings of anticipatory shifts in the receptive fields of neurons in parietal cortex⁷ and superior colliculi⁸.

For most experiments, observers made 20° left-to-right saccades in a dimly lit room from a fixation point F_0 (at -10°) on an otherwise featureless red screen, to a target F_1 presented at $+10^\circ$ after a ready signal. Green equiluminant vertical bars were briefly flashed at various positions, and observers reported their location with reference to a ruler that appeared on the screen shortly after the end of each saccade. The results in Fig. 1 show that bars displayed at physical positions of either 0 or -20° (squares and triangles, respectively) were systematically mislocalized in the direction of the saccade. The shift effects begin 50 ms before the saccade, rising to a maximum (about 10°) in a critical period just before the saccade onset ($-25 < t < 0$ ms). After the saccade had finished, localization of the bar was again veridical, although it now fell on a different retinal location from before. The results were quite different for bars displayed to the right of the target F_1 at $+20^\circ$ (circles in Fig. 1). Here the apparent position was displaced against the direction of saccades before the eye movement. The apparent length of the long bars (50°) did not change significantly under any condition (data not shown). The continuous curves in Fig. 1 and in all other figures come from the simple model described in Methods, which assumes a change in the origin of perceptual space from F_0 to F_1 and a perceptual compression.

Figure 2 describes the pattern of results found within the critical period, -25 to 0 ms (relative to saccade onset), for bars displayed over a wide range of spatial positions. Bars displayed to the left of F_1 were displaced in the direction of the saccade, whereas those to the right were displaced in the opposite direction, with a tendency for the data to cluster around F_0 and F_1 (dotted lines). Bars falling over a wide range (-5 to 30°) were mostly perceived near F_1 , the saccade target. A similar pattern of results was observed for right-left, vertical and smaller (10°) saccades. These findings imply a compression of space within the critical interval. To confirm that the compression was real, we measured the vernier displacement of two half bars, first when spatially separated but displayed simultaneously and briefly. In this situation, the individual upper and lower bars behaved exactly like the full bars, resulting in the impression of

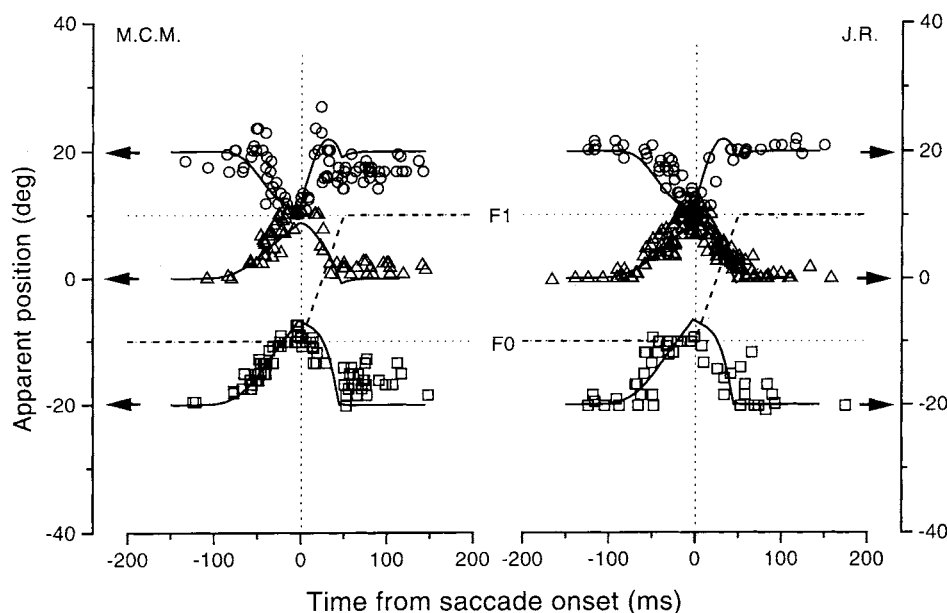


Figure 1 Apparent position of a vertical bar briefly flashed at positions -20° (squares), 0° (triangles) or $+20^\circ$ (circles), as a function of time relative to saccade onset. Each point represents a single observation. Saccades from F_0 to F_1 are indicated by the dashed line. The solid curves are simulations from the model described in Methods. Similar results have been obtained with 6 other observers, all initially naive of the aims of the experiment.

a collinear bar when the two halves were presented within the critical period, one at 0 and the other at $+20^\circ$.

We also measured the apparent displacement of two half bars displayed collinearly but at different times, one 75 ms before the other. Subjects either reported the apparent separation of the bars (open circles in Fig. 3), or annulled it with a physical offset (filled squares: see Methods). The most interesting period is between -100 and -75 ms, when both bars were displayed in the same position to stationary eyes, but only the later bar fell just before saccade onset (when the mislocalization is strongest). When the spatial position of the bars was near 0° , the later bar was mislocalized with respect to the earlier in the direction of the saccade (positive offset for the observer). However, for bars displayed at 20° the effect inverted, showing a shift against the direction of the saccade. The model shown by continuous lines fits the results reasonably well, although the data measured at 20° eccentricity are more noisy.

For a stronger test of compression of visual space, we displayed up to four bars within a 20° region centred on F_1 and asked observers to report how many they saw. Well before or after the saccade, observers seldom erred, but near the saccade onset they usually reported seeing only one bar, no matter how many were displayed (Fig. 4). However, if the bars differed in colour or orientation, the

observers could perceive them as separate items, grouped closely together. We also displayed briefly several natural scenes such as that shown in Fig. 4 (lower left). The image on the right shows the deformation predicted by the model of equation (1). Most observers, both trained and naive, agreed that the deformed image resembled closely what they saw when the bridge scene was presented just before saccade onset. No compression of either bar position or natural scene was observed during 'stimulated saccades' (created by viewing the display through a rotatable mirror displacing the image 20° at saccadic speeds).

Our results imply a bidirectional compression of space that cannot be explained by a simple translation of coordinates^{1-3,9}; this could cause a uniform change in apparent spatial position, but not position-dependent mislocalization of single targets, nor gross misjudgements of relative position and of number of display items, and deformation of natural scenes. The presaccadic mislocalization against the saccade is particularly surprising, and directly contradicts some previous studies¹⁰ that report displacement in the saccadic direction at all positions. There could be many reasons for the difference, including the fact that many experiments were done in the dark¹⁰. The use of equiluminant stimuli could also be important in bringing forward in time the effects of mislocalization

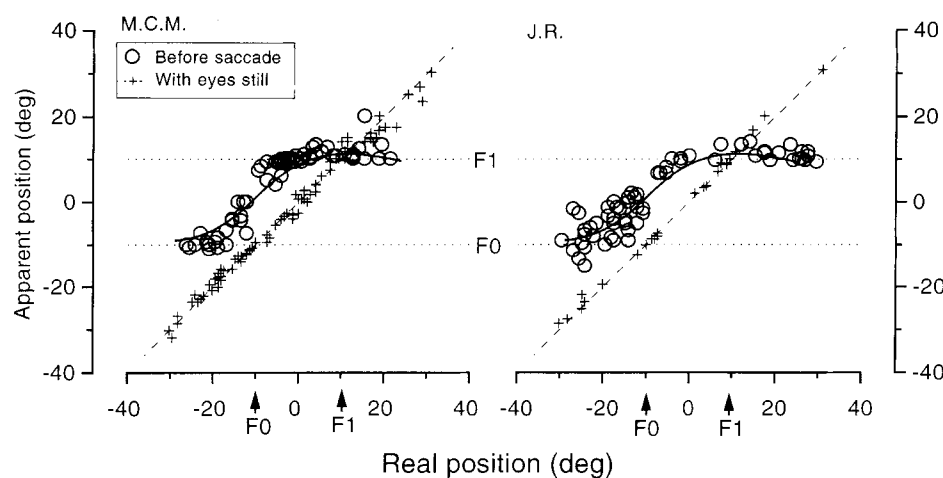


Figure 2 Apparent position of a green equiluminant bar present just before the saccade onset (-25 to 0 ms) as a function of real position (circles) as subjects M.C.M. and J.R. saccaded left to right. To avoid response stereotyping, the bars were presented randomly with various latencies, but only data with appropriate delays are plotted here. The crosses refer to measurements made for the same briefly exposed bars with constant fixation at F_0 , demonstrating that judgements are veridical with normal viewing ($r > 0.995$ for both observers). The dashed line represents veridical judgement, and the solid line the simulation from the model described in Methods. This pattern of results has been verified with right to left and vertical saccades, and on 8 naive observers.

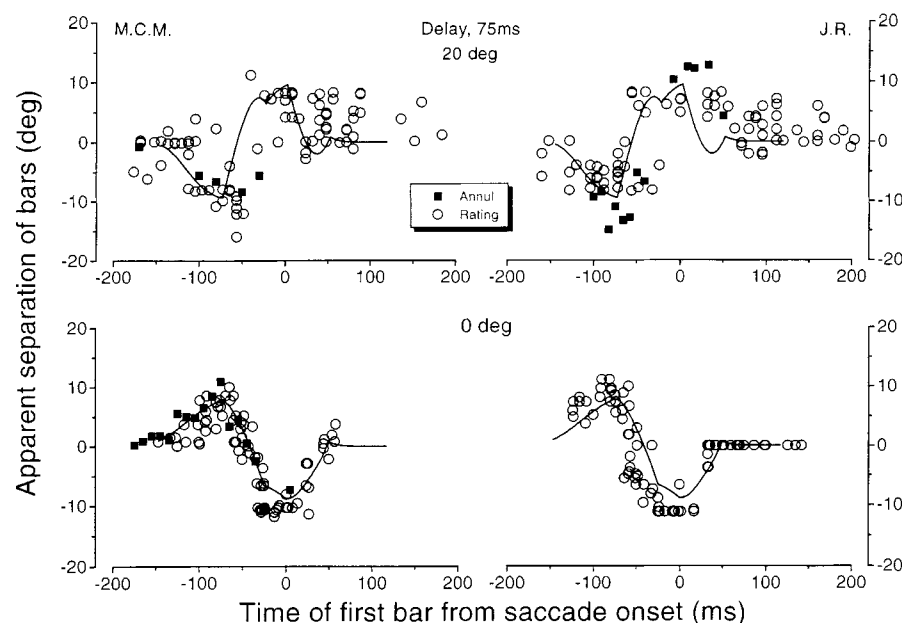


Figure 3 Vernier judgement; or annulment of two collinear half bars. The top half bar was briefly displayed 75 ms before or after the bottom half, both at position of 0 or 20° . Open circles show observers' reports of the apparent size and direction of the displacement, and filled symbols the estimate obtained by the annulling technique described in Methods. For delays between -100 and -75 ms (when both bars were displayed to the same position to stationary eyes), there was a positive displacement (meaning that the later bar was displaced rightwards) when the bars were displayed near 0° , and a negative displacement when the bars were displayed at 20° . The trend later reversed in both cases to a maximum in the opposite direction at $t = 0$ ms. Smooth curves through the data are predictions of the model.

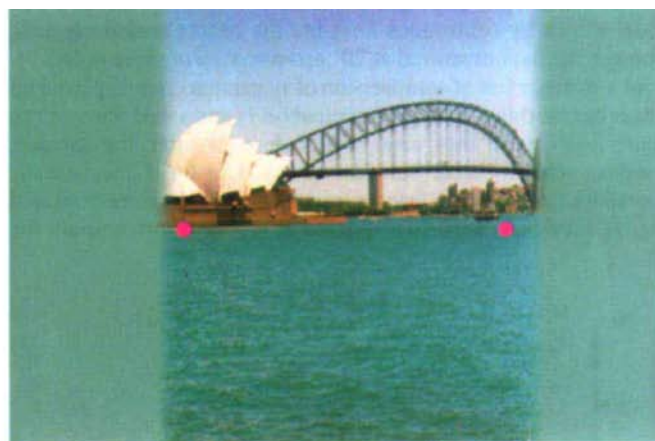
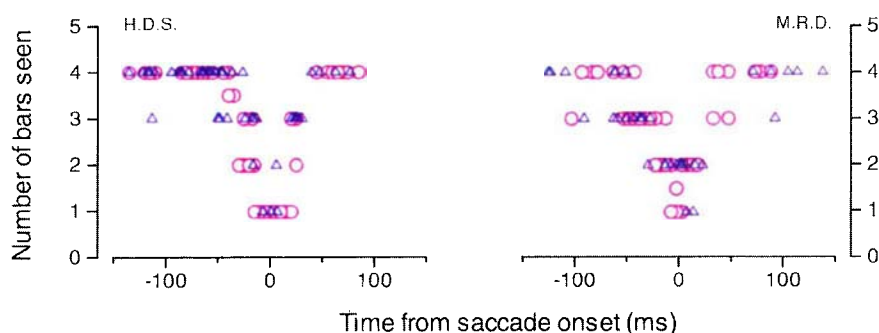


Figure 4 Top, Multiple bars (between 0 and 4) were assigned at random to 4 positions straddling F_1 (0, 7, 14 and 20°), and briefly displayed simultaneously at random delays with respect to saccadic onset. Observers H.D.S. and M.R.D. never reported a bar when none had been displayed, and always reported accurately when only 1 had been displayed. The graphs show the results only for when 4 bars were actually displayed: errors occur when the bars are displayed near the onset of the saccade. Circles, equiluminant green bars; triangles, black bars. Bottom, Natural scenes such as that on the left were briefly displayed to 13

observers (most naive to the aims of the test) at various intervals relative to a saccade. When the scene was displayed just before the saccade ($-25 \text{ ms} < t < 0 \text{ ms}$), it appeared compressed, similar to the simulation shown on the right, produced by applying equation (1) with $t = -20 \text{ ms}$. Some subjects (two) saw the bridge as symmetrical, others (two) reported that the Opera House sometimes detached itself from the bridge, but all observers saw the scene compressed.

(as they have longer processing latencies¹¹). However, evidence for position-dependent mislocalization does exist^{5,12–15}, although none of the earlier studies highlighted the effects with the use of multiple rather than single stimuli, or tested compression in other ways. Here the evidence for compression is unmistakable, and not dismissible as artefact⁵. None of our results, such as seeing four widely separated bars as one, can be attributed to saccadic suppression, because we used equiluminant stimulus, no less visible in saccades than in normal viewing¹⁶, and compression does not occur during 'simulated saccades' when the scene is moved rapidly by mechanical means.

We have constructed a two-component model for our data: a presaccadic shift of the same direction and amplitude of the saccade, together with a compression. The shift fits well with the anticipatory behaviour of cells in the lateral intraparietal cortex (LIP) and superior colliculi^{7,8}. Many of the cells in these areas respond presaccadically to stimuli that will fall onto their receptive fields after the saccade has been completed, making them likely candidates to mediate the shift in visual coordinates that must accompany each saccade. The physiological substrate for the compression is not so clear, but the receptive fields of some LIP cells may change shape before saccades¹⁷ (often expanding in size), consistent with compression. It may also be relevant that the magnocellular pathway, which provides the major input to parietal cortical areas important for visual localization¹⁸, seems to be selectively suppressed during saccades¹⁶, and the suppression could lead to a distortion in the metric of the internal representation of space.

Whatever the neural substrate for the compression, its functional role may be to increase tolerance to mismatches between the intention-to-move signal and the actual displacement, possibly explaining why observers are so poor in detecting image displacement during saccades¹⁹. When stimuli are occluded for the critical period around the saccade onset, and so are not subjected to compression, detection of image displacements improves greatly²⁰. The compression of space may reflect a moment of maximal plasticity and uncertainty, when objects are compressed within a restricted region, to be relocated into their real positions only after the new coordinate system has been fully established, taking into account new visual information. This could allow for a remapping of the effects of eye movements with only two coarse signals, and without recourse to expensive memory maps of the external world. □

Methods

General. Observers, with normal or corrected acuity and normal oculomotility, fixated binocularly a point F_0 at -10° (left of screen centre) on an otherwise featureless red screen ($70 \times 50^\circ$ at 25 cm, mean luminance 17 cd m^{-2}), surrounded by 2-m-square white card, illuminated to similar luminance and colour. After a warning signal, they made a saccade to a target F_1 which appeared at $+10^\circ$ and remained on until the next trial. The stimulus (usually a green bar subtending $1.5 \times 50^\circ$) was presented for 1 frame (8 ms) to a random position after a random interval. For the experiments in Figs 1 and 2, observers reported verbally the apparent position of the centre of the bar with respect to a ruler presented one second later (other experiments are described

in the figure legends). The stimuli were generated on a Mitsubishi colour monitor at 120 Hz by a computer-controlled framestore (Cambridge Research Systems VSG). Horizontal eye movements were monitored by an infra-red eye tracker (HVS Image Ltd), and actual delay of target presentation (with respect to the saccade onset) were determined after each trial by computer. For the results shown in Figs 1–3, the bar was equiluminant with the background (as these stimuli are not suppressed during saccades¹⁶); the same pattern of results has been obtained with light or dark bars.

Vernier annulment. For the vernier annulment task, the top half of the bar was briefly displayed 75 ms before or after the bottom half, and physically misaligned. Observers reported whether the top half appeared displaced to the left or right. Adaptive routines adjusted the physical offsets at each display time, based on the previous responses. Final thresholds were estimated by fitting a cumulative gaussian to the probability-of-seeing curves, comprising 80–250 data points. The standard deviation of the gaussian which gives an estimate of error, was on average 0.7°, and did not vary systematically with display time.

Model. To simulate the data of Figs 1–4, we assume that apparent visual direction P of any point x in external space is given by:

$$P(x, t) = E(x, t) \times C(x, t) + O(t) \quad (1)$$

$E(x, t)$ is retinal eccentricity (the difference between physical external space and eye position); $C(x, t)$ is a compression function weighting the eccentricities; $O(t)$ is an 'extra-retinal' signal, that can be considered to set the origin of the internal representation of space, usually given by eye position (ignoring head movements for now). On each saccade, $O(t)$ changes gradually, following the sum of the fade-out and fade-in functions that weight the present and future eye position:

$$O(t) = F_0 \left(1 - \int_0^t e^{-\frac{(t-\tau)^2}{2\sigma_0^2}} d\tau \right) + F_1 \int_0^t e^{-\frac{(t-\tau)^2}{2\tau_1^2}} d\tau$$

The function $C(x, t)$ is given by

$$C(x, t) = e^{-k \left[E(x, t) / (O(t) - Y(t)) / (F_0 - F_1) \right]^2}$$

where k is a constant and $Y(t)$ is eye position. $C(x, t)$ bears a similarity to the inverse of the cortical magnification function²² (and may reflect a visual process designed to compensate for it). The free parameters in the equations were adjusted to achieve best fits to all the data for both subjects: $\sigma_0 = 30$ ms, $\sigma_1 = 38$ ms, $\tau_0 = 20$ ms, $\tau_1 = -40$ ms, $k = 1.48$, $\beta = 1.35$.

Received 7 June 1996; accepted 29 January 1997.

1. von Helmholtz, H. *Handbuch der Physiologische Optik* (1866); translated by Southall, J. P. C. *Treatise on Physiological Optics* (Dover, New York, 1963).
2. Sperry, R. W. Neural basis of the spontaneous optokinetic response produced by visual inversion. *J. Comp. Physiol. Psychol.* **43**, 482–489 (1950).
3. Von Holst, E. & Mittelstaedt, H. Das Reafferenzprinzip. *Naturwissenschaften* **37**, 464–476 (1954).
4. Mackay, D. M. Elevation of visual threshold by displacement of visual images. *Nature* **225**, 90–92 (1970).
5. O'Regan, J. K. Retinal versus extraretinal influences in flash localization during saccadic eye movements in the presence of a visible background. *Percept. Psychophys.* **36**, 1–14 (1984).
6. Sperling, G. in *Eye movements and their role in visual and cognitive processes* (ed. Knowlton, E.) 307–351 (Elsevier, Amsterdam, 1990).
7. Duhamel, J.-R., Golby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).
8. Walker, M. F., Fitzgibbon, J. & Goldberg, M. E. Neurons of the monkey superior colliculus predict the visual result of impending saccadic eye movements. *J. Neurophysiol.* **73**, 1988–2003 (1995).
9. Matin, L. in *Handbook of Sensory Physiology VII/4: Visual Psychophysics* (eds Jameson, D. & Hurvich, L. M.) 331–380 (Springer, Berlin, 1972).
10. Dassonville, P., Schlag, J. & Schlag-Rey, M. Oculomotor localization relies on a damped representation of saccadic eye movement displacement in human and nonhuman primates. *Visual Neurosci.* **9**, 261–269 (1992).
11. Bowen, R. W. Latencies for chromatic and achromatic visual mechanisms. *Vision Res.* **21**, 1457–1466 (1981).
12. Matin, L. & Pearce, D. G. Visual perception of direction for stimuli flashed during voluntary saccadic eye movements. *Science* **148**, 1485–1487 (1965).
13. Bischof, N. & Kramer, E. Untersuchungen und Überlegungen zur Richtungswahrnehmung bei willkürlichen sakkadischen Augenbewegungen. *Psychol. Forsch.* **32**, 185–218 (1968).
14. Honda, H. Saccade-contingent displacement of the apparent position of visual stimuli flashed on a dimly illuminated structured background. *Vision Res.* **33**, 709–716 (1993).
15. Honda, H. Visual mislocalization produced by a rapid image displacement on the retina: examination by means of dichoptic presentation of a target and its background. *Vision Res.* **35**, 3021–3028 (1995).
16. Burr, D. C., Morrone, M. C. & Ross, J. Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature* **371**, 511–513 (1994).
17. Ben Hamed, S. & Duhamel, J.-R., Bremmer, F. & Graf, W. Dynamic changes in visual receptive field organization in the macaque lateral intraparietal area (LIP) during saccade preparation. *Soc. Neurosci. Abstr.* Part 2, 1619 (1996).
18. Mishkin, M., Ungerleider, L. G. & Macko, K. A. Object vision and spatial vision: two cortical pathways. *Trends Neurosci.* **6**, 414–417 (1983).

19. Bridgeman, B. & Stark, L. Ocular proprioception and efference copy in registering visual direction. *Vision Res.* **31**, 1903–1913 (1991).
20. Deubel, H., Schneider, W. X. & Bridgeman, B. Postsaccadic target blanking prevents saccadic suppression of image displacement. *Vision Res.* **36**, 985–996 (1996).
21. Bridgeman, B., Van der Heijden, A. H. C. & Velichkovsky, B. M. A theory of visual stability across saccadic eye movements. *Behav. Brain Sci.* **17**, 247–292 (1994).
22. Daniel, P. M. & Whitteridge, D. The presentation of the visual field on the cerebral cortex in monkeys. *J. Physiol. (Lond.)* **159**, 203–221.

Acknowledgements. This work was supported by the Australian ARC and the Italian CNR. We thank Stephen Peter for permission to use the photograph in Fig. 4.

Correspondence and requests for materials should be addressed to D.C.B. in Pisa (e-mail: dave@neuro.in.pi.cnr.it).

Perceived geometrical relationships affected by eye-movement signals

Rick H. Cai*, Alexandre Pouget†, Madeleine Schlag-Rey* & John Schlag*

* Brain Research Institute, UCLA School of Medicine, Los Angeles, California 90095, USA

† Institute for Cognitive and Computational Sciences, Georgetown University, Washington DC 20007, USA

To determine the location of visual objects relative to the observer, the visual system must take account not only of the location of the stimulus on the retina, but also of the direction of gaze¹. In contrast, the perceived spatial relationship between visual stimuli is normally assumed to depend on retinal information alone, and not to require information about eye position. We now show, however, that the perceived alignment of three dots—tested by a vernier alignment task^{2,3}—is systematically altered in the period immediately preceding a saccade. Thus, information about eye position can modify not only the perceived relationship of the entire retinal image to the observer, but also the relations between elements within the image. The processing of relative position and of egocentric (observer-centred) position may therefore be less distinct than previously believed^{4–6}.

Our perception of a stable visual world is achieved, in part, by computing the egocentric position of stimuli through summing their retinotopic location with an internal representation of the direction of the eye⁶. However, this process is not always perfect and can sometimes lead to mislocalization. Hence, if a flash is presented shortly before a saccade, owing to the large delay in the afferent visual pathway (up to 100 ms), the internal representation of the eye position may have already started to change in the saccade direction by the time the visual signal reaches the cortex. The spatial location of the visual stimulus would then be miscalculated⁷. Indeed, it is known that a single point of light flashed within a critical period (about 100 ms) before a saccade is perceived as displaced in the direction of the saccade (a phenomenon called presaccadic mislocalization)^{7–10}. It has also been shown that a continuously lit spot of light, starting before the critical period and extinguished before saccade onset, is not mislocalized, presumably because its stable location has already been established¹⁰.

We investigated whether the perception of the relative position of nearby dots can be affected by the imminence of a saccade. We used a 3-dot vernier test in which the lower and upper dots were continuously lit until the presentation of the middle dot (for 4 ms), and then all three dots were extinguished together (Fig. 1, and see Methods). We carried out this test during the latency period of a visually evoked saccade (saccade trial) and compared the results with those obtained in the absence of an impending saccade (the

subject kept fixating). In each no-saccade trial, the positions and the timing of all visual stimuli duplicated those of a saccade trial (matched trials). The eye position was monitored by an infrared eye tracker. Only those saccade trials in which all visual stimuli were presented and extinguished before the eyes started to move were compared with no saccade trials. This ensured that the same visual stimuli were delivered onto identical retinal spots across conditions. Consequently, any difference of judgement must be of extraretinal origin.

When several visual stimuli are concurrently on the retina, the retinal information alone suffices to specify completely their relative positions and, as far as we know, in previous studies there has been no indication that extraretinal factors may be involved in such a situation. A distortion of vernier alignment by presaccadic mislocalization would indicate, however, that the perception of relative position is also dependent upon an extraretinal factor, even though the latter is in principle not needed.

Five human subjects (4 naive, 1 author) with normal or corrected vision were tested. For stimuli presented within 100 ms before a saccade, four of the five subjects showed large bias shifts in their vernier curves from the no-saccade positions (shifts ranging from 52' to 59' before 4° rightward saccades; Fig. 2). The fifth naive subject (A.K.) had a smaller shift (42'). The shift was such that the middle dot always appeared to be displaced in the direction of the saccade. Two of the subjects were tested further in a complementary situation, in which the display was to the left and 4° leftward saccades were made. Again, the middle dot was displaced in the same direction as the saccade (63' and 45'). Thus, in all cases the direction of the shift was consistent with the direction of the presaccadic mislocalization observed previously⁷⁻¹⁰ (except in a recent study conducted under very different experimental conditions¹¹). Vernier acuity (slope) varied across subjects, but differed little between saccade and no-saccade conditions within the same subject.

Could the observed shifts be caused by saccadic suppression¹², rather than presaccadic mislocalization? This seems unlikely as saccadic suppression would predict shallower vernier curves but not a systematic directional shift in the curves. Another explanation would attribute our results to a covert attention shift before the saccade. If this were correct, the same shift would occur with or

without an overt saccade¹³; however, our no-saccade condition, used as a baseline, has demonstrated the dependence of the shifts on the occurrence of overt saccades.

To verify further that the observed vernier shifts were caused by presaccadic mislocalization, we did two additional experiments. In the first, we flashed all three dots at the same time shortly before the saccade. If the observed vernier shift is indeed due to presaccadic mislocalization, it should now disappear, because all three dots should be mislocalized by the same amount. This prediction was tested in three subjects (Fig. 3a): for all three, the shift before saccade disappeared completely. In the second experiment, we exploited the finding that under our conditions the presaccadic mislocalization begins at ~100 ms before the saccade onset⁸⁻¹⁰. If the observed vernier shift is due to presaccadic mislocalization, the shift should not exist before this time. We tested this prediction in four subjects. The experiment was similar to that shown in Fig. 2, but instead of flashing the middle dot very close to saccade onset, it was flashed (and subsequently all three dots extinguished) more than 100 ms before saccade onset. As predicted, there was no systematic vernier curve shift between the saccade and no-saccade trials under this condition (Fig. 3b-d).

These two experiments showed that an impending saccade, by itself, is not sufficient to create a vernier curve shift. A specific set of stimulus characteristics and timing conditions, consistent with the

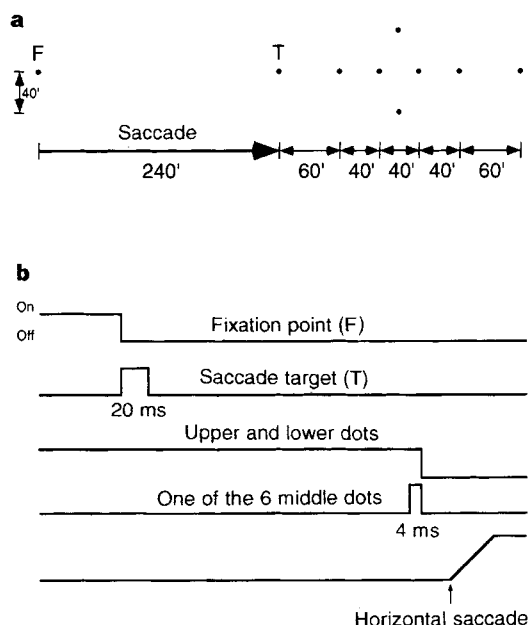


Figure 1 a, Stimulus locations drawn to scale (for all experiments). b, Stimulus timing (for Figs 2 and 3b, c).

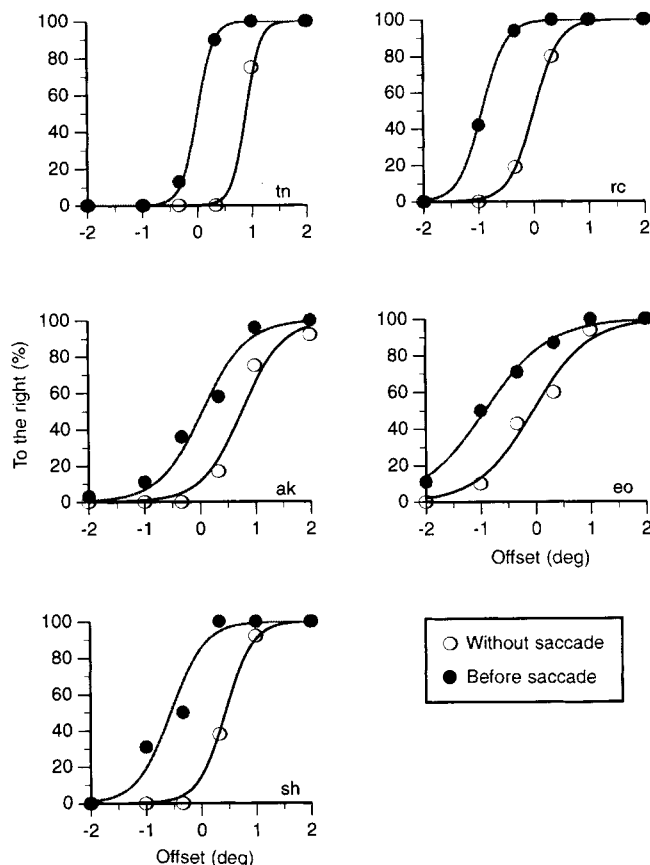


Figure 2 Shift of vernier curves within 100 ms before 4° rightward saccades for all five subjects. Abscissa: horizontal offset of the middle dot from the lower and upper dots (+ is to the right). Ordinate: percentage of seeing the middle dot to the right of the lower and upper dots. The lower and upper dots are lit from the beginning of the trial, the middle dot is flashed (Fig. 1b). The saccade curve is based on trials in which the middle dot flashed within 100 ms before the saccade onset. The saccade and the no-saccade curves are based on matched trials (see text). Constant errors (non-zero bias of no-saccade curves) exist in three subjects, as have been observed elsewhere²⁹. The vernier curve shift is consistent across subjects, regardless of how much constant error each subject has.

mislocalization explanation, is also required for the shift to occur. These results also rule out the possibility that the vernier curve shift could be due to a motor response bias, conceivably created by imminent saccades always in the same direction.

Relative position judgement has been studied with both foveal^{14,15} and extrafoveal stimuli^{16,17}. In both cases, the subjects' performance has been explained by mechanisms operating at early stages of visual processing^{16,17}. Most electrophysiological studies in this area have also been centred on the retinotopic neuronal responses in structures such as the primary visual cortex (V1) and the lateral geniculate nucleus^{18–20}. The alignment of visual stimuli can be biased by other retinal cues, however, particularly those related to image motion^{21,22}. Such results raise doubts about the exclusive

dependence of vernier alignment upon the responses of striate neurons²², but they agree with the idea that the perception of vernier alignment is entirely derived from the retinal input, because only visual variables were manipulated in these experiments. Here, by contrast, without any change in the visual input, we have shown that an extraretinal factor suffices to alter the perception of vernier alignment. To our knowledge, previous results never indicated that the perception of relative position could depend on neurons combining retinal and extraretinal signals. One possibility is that the calculation of vernier alignment is not completed—or at least the final decision not made—until a higher cortical area is reached, such as the parietal cortex, where a large percentage of visually responsive neurons are modulated by the eye position²³, and where the receptive fields of some neurons are shifted shortly before a saccade²⁴. Alternatively, an extraretinal signal may enter early visual areas, altering the retinotopic responses of visual cells. Neurons modulated by eye position have been reported in V1 (refs 5 and 25) and V3A (ref. 4), and modelling studies have suggested that such early modulation can perform a spatial-processing function similar to that found in the parietal cortex^{26,27}. Both of these possibilities suggest that the processing of relative position and the processing of egocentric position are more closely related than was previously thought^{4–6}. □

Methods

The subject sat in total darkness with the head immobilized by a bite plate, 130 cm away from the visual display. The eye movements were recorded with an Ober2 infrared eye-orbit scanner. The viewing was monocular (right eye). All visual stimuli were small green light-emitting diodes (LEDs) ($r = 0.06$ deg). The top and bottom dots were vertically aligned and served as reference points; the middle dot took one of six possible horizontal positions (Fig. 1). The luminance of the outer dots was 100 cd m^{-2} , and the middle dot (flashed) was adjusted to have the same subjective brightness. At the beginning of a saccade trial, a central fixation point (straight ahead) and the upper and lower reference dots (at 6° to the right) were turned on. After the computer determined that the eyes had been steadily on the fixation point for a duration that was randomly varied between 700 and 1,800 ms (to prevent anticipatory saccades), the fixation point was turned off and the saccade target was flashed for 20 ms at 4° to the right. The subject was to saccade towards this target as soon as possible (average saccade latency was ~ 200 ms). The saccade onset time was predicted from the average saccade latency of the past 4 trials. Some time before this predicted onset (20–30 ms for Figs 2 and 3a; 100–200 ms for Fig. 3b–d), the middle dot was flashed for 4 ms. All three dots were then extinguished at the same time. This method produced an average 88.2% of saccade trials in which all visual stimuli were presented before the saccade onset (Fig. 2). After the saccade, the subject reported, by pressing one of the two buttons, whether the middle dot appeared to the left or right of the upper and lower dots. In no-saccade trials, the exact same sequence of visual stimuli was displayed while the subject was instructed to keep fixating. Saccade and no-saccade trials were run in separate blocks of 200 trials per block. Only one block was run on a given day to prevent fatigue.

The vernier display was not placed at the site of fixation because the subjects found it very difficult to pay attention to the foveal vernier test, while at the same time attempting to move the eyes quickly toward a peripheral saccade target. Thus, the vernier display was placed close to the saccade target. The larger the saccade, the greater the magnitude of mislocalization. But as the vernier display needs to be placed close to the saccade target, a larger saccade would also mean a more peripheral vernier display, and hence poorer vernier acuity. After repeated trials, we found that placing the display at 6° best satisfied the requirements of the experiment. The optimum dot separation depends on eccentricity¹⁶ and we found $40'$ to be a good choice for our specific test conditions.

Received 3 March; accepted 14 March 1997

- Grüsser, O.-J. Space perception and the gazemotor system. *Hum. Neurobiol.* **1**, 73–76 (1982).
- Westheimer, G. Visual acuity and hyperacuity. *Invest. Ophthalmol.* **14**, 570–572 (1975).
- Levi, D. M. & Klein, S. A. Limitations on position coding imposed by undersampling and univariance. *Vision Res.* **36**, 2111–2120 (1996).

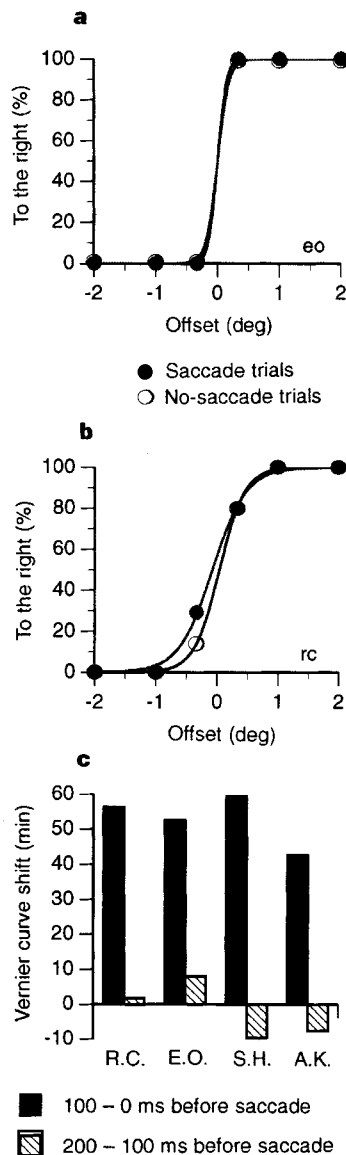


Figure 3 **a**, Lack of vernier curve shift when all three dots simultaneously flashed within 100 ms before 4° rightward saccades (two curves overlap). One subject's data are shown here (all three subjects' data similar). The vernier curve is steeper than that in Fig. 2. This is consistent with the finding that the highest vernier acuity is achieved with synchronous stimulus onset and offset²⁸. (The precise slope is beyond the resolution limit of our display apparatus.) **b**, Vernier curve shifted little when the middle dot flashed between 200 and 100 ms before saccade onset (subject R.C.). **c**, Amount of vernier curve shift (shift in point of subjective equality) in early and late latency period for all four subjects tested. Same experimental conditions as for Fig. 2.

4. Galletti, C. & Battaglini, P. P. Gaze-dependent visual neurons in area V3A of monkey prestriate cortex. *J. Neurosci.* **9**, 1112–1125 (1989).
5. Weyand, T. G. & Malpeli, J. C. Responses of neurons in primary visual cortex are modulated by eye position. *J. Neurophysiol.* **69**, 2258–2260 (1993).
6. MacKay, D. M. in *Handbook of Sensory Physiology* (ed. Jung, R.) VII/3, 307–332 (Springer-Verlag, Berlin, 1973).
7. Matin, L. in *Handbook of Sensory Physiology* (eds Jameson, D. & Hurvich, L.) VII/4, 307–332 (Springer-Verlag, Berlin, 1972).
8. Honda, H. in *Attention and Performance* (ed. Jeannerod, M.) XIII, 567–582 (Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1990).
9. Honda, H. The time courses of visual mislocalization and of extraretinal eye position signals at the time of vertical saccades. *Vision Res.* **31**, 1915–1921 (1991).
10. Schlag, J. & Schlag-Rey, M. Illusory localization of stimuli flashed in the dark before saccades. *Vision Res.* **35**, 2347–2357 (1995).
11. Ross, J., Morrone, M. C. & Burr, D. *Nature* **386**, 598–601 (1997).
12. Volkman, F. C., Schick, A. M. L. & Riggs, L. A. Time course of visual inhibition during voluntary saccades. *J. Opt. Soc. Am.* **58**, 562–569 (1969).
13. Posner, M. I. Orienting of attention. *Quart. J. Exp. Psych.* **32**, 3–25 (1980).
14. Westheimer, G. & McKee, S. P. Spatial configurations for visual hyperacuity. *vision Res.* **17**, 941–947 (1977).
15. Groll, S. L. & Hirsch, J. Two-dot vernier discrimination within 2.0 degrees of the foveal center. *J. Opt. Soc. Am. A*, **4**, 1535–1542 (1987).
16. Whitaker, D., Rovamo, J., MacVeigh, D. & Makela, P. Spatial scaling of vernier acuity tasks. *Vision Res.* **32**, 1481–1491 (1992).
17. Levi, D. M. & Waugh, S. A. Spatial scale shifts in peripheral vernier acuity. *Vision Res.* **34**, 2215–2238 (1994).
18. Parker, A. & Hawken, M. Capabilities of monkey cortical cells in spatial-resolution tasks. *J. Opt. Soc. Am.* **2**, 1101–1114 (1985).
19. Swindale, N. V. & Cynader, M. S. Vernier acuity of neurons in cat visual cortex. *Nature* **319**, 591–593 (1986).
20. Shapley, R. & Victor, J. Hyperacuity in cat retinal ganglion cells. *Science* **231**, 999–1002 (1986).
21. Fahle, M. & Poggio, T. Visual hyperacuity: spatiotemporal interpolation in human vision. *Proc. R. Soc. Lond. B* **213**, 451–477 (1981).
22. De Valois, R. L. & De Valois, K. K. Vernier acuity with stationary moving Gabors. *Vision Res.* **31**, 1619–1626 (1991).
23. Andersen, R. A., Essick, G. K. & Siegel, R. M. Encoding of spatial location by posterior parietal neurons. *Science* **230**, 456–458 (1985).
24. Duhamel, J., Colby, C. L. & Goldberg, M. E. The updating of the representation of visual space in parietal cortex by intended eye movements. *Science* **255**, 90–92 (1992).
25. Trotter, Y., Celebrini, S., Stricane, B., Thorpe, S. & Imbert, M. Modulation of neural stereoscopic processing in primate area V1 by the viewing distance. *Science* **257**, 1279–1281 (1992).
26. Pouget, A., Fisher, S. A. & Sejnowski, T. J. Egocentric representation in early vision. *J. Cogn. Neurosci.* **5**, 150–161 (1993).
27. Pouget, A. & Sejnowski, T. J. Spatial transformations in the parietal cortex using basis functions. *J. Cogn. Neurosci.* **9**, 222–237 (1997).
28. Wehrhahn, C. & Westheimer, G. Temporal asynchrony interferes with vernier acuity. *Vis. Neurosci.* **10**, 13–19 (1993).
29. Lindblom, B. & Westheimer, G. Binocular summation of hyperacuity tasks. *J. Opt. Soc. Am. A* **6**, 585–589 (1989).

Acknowledgements. We thank D. Bavelier and J. Thomas for their comments and valuable help. This work is supported by the US National Eye Institute.

Correspondence and requests for materials should be addressed to R.C. (e-mail: rc@ucla.edu).

Temporal dynamics of brain activation during a working memory task

**Jonathan D. Cohen[†], William M. Perlstein[†],
Todd S. Braver[†], Leigh E. Nystrom[†],
Douglas C. Noll[‡], John Jonides[§] &
Edward E. Smith^{*}**

^{*} Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

[†] Department of Psychiatry, University of Pittsburgh, 3811 O'Hara Street, Pittsburgh, Pennsylvania 15213, USA

[‡] Department of Radiology, University of Pittsburgh Medical Center, 200 Lothrop Street, Pittsburgh, Pennsylvania 15213, USA

[§] Department of Psychology, University of Michigan, 525 East University, Ann Arbor, Michigan 48109, USA

Working memory is responsible for the short-term storage and online manipulation of information necessary for higher cognitive functions, such as language, planning and problem-solving^{1,2}. Traditionally, working memory has been divided into two types of processes: executive control (governing the encoding manipulation and retrieval of information in working memory) and active maintenance (keeping information available 'online'). It has also been proposed that these two types of processes may be subserved

by distinct cortical structures, with the prefrontal cortex housing the executive control processes, and more posterior regions housing the content-specific buffers (for example verbal versus visuospatial) responsible for active maintenance^{3,4}. However, studies in non-human primates suggest that dorsolateral regions of the prefrontal cortex may also be involved in active maintenance^{5–8}. We have used functional magnetic resonance imaging to examine brain activation in human subjects during performance of a working memory task. We used the temporal resolution of this technique to examine the dynamics of regional activation, and to show that prefrontal cortex along with parietal cortex appears to play a role in active maintenance.

Neurologically normal subjects (5 males, 5 females; ages 18–34) were scanned while performing a sequential-letter memory task (Fig. 1). This task has reliably produced activation of cortical regions that are believed to be involved in working memory^{9,10}. Memory load was varied parametrically to identify these regions sensitively¹¹. In addition, the rate of stimulus presentation was slowed in order to acquire multiple scans during each trial, and thereby track the dynamics of activation (Fig. 2). We reasoned that temporal information, together with the manipulation of memory load, would provide new information permitting a finer analysis of the cognitive functions associated with activated regions than has previously been possible. Specifically, we predicted that sensory and motor processes (ones not involved in working memory) would exhibit transient increases in activation associated with stimulus presentation and response execution (peaking after a delay of about 5 s, as a result of the well-characterized lag in haemodynamic response^{12–14}), but should not vary as a function of memory load. We predicted that the areas involved in working memory would vary as a function of memory load, with greater activation at higher levels of load. Furthermore, we predicted that such load-sensitive areas would dissociate into two types: those involved in active maintenance would exhibit sustained activation throughout the trial, whereas those involved in other working memory functions (assumed to be time-limited, such as updating working memory contents, decision processes and so on) would exhibit transient activation (such as sensory and motor processes) but would peak higher (or last longer) at higher levels of load. Thus these areas would show an interaction between the effects of time and load.

Our findings, from pooled data for 10 right-handed subjects, reveal each of the patterns predicted above (Fig. 3). As expected, regions within visual, motor and somatosensory cortex all exhibit strong effects of time, but no effect of memory load (see Table 1 and Fig. 3a). Motor and somatosensory regions are left-lateralized, consistent with right-handed response. The time course of activation in all of these regions concurs with other studies focusing specifically on these systems^{12,13} and validates our ability to track the dynamics of activation using this method.

The distribution of regions showing sensitivity to memory load corresponds well with previous observations using this task^{10,11} and with structures thought to be involved in working memory. These include dorsolateral prefrontal cortex (PFC), more posterior and inferior regions of frontal cortex (including Broca's area), and posterior parietal cortex. As predicted, however, two different temporal patterns are evident among these regions. Within anterior frontal cortex, including dorsolateral PFC (BA46/9), only regions showing an effect of load are observed but none showing an interaction with time (Table 1 and Fig. 3b). Such regions are also observed within more posterior structures, including Broca's area (BA44) and posterior parietal cortex (BA40), but in posterior areas they co-occur with (and sometimes are directly adjacent to) other regions that show an interaction between load and time (Fig. 3c).

The pattern of activation observed within dorsolateral PFC (greater with higher levels of load, and sustained throughout the trial) is consistent with a role in the active maintenance of information in working memory. This suggests that PFC is not exclusively

involved in transient processes, such as assigning temporal order, updating the contents of working memory, or other memory-related processes (such as encoding or retrieval from longer-term stores¹⁵). Previous neuroimaging studies have found PFC activation during the retention interval of a working memory task, but lacked the temporal resolution to determine whether such activation was sustained throughout the entire retention interval, or occurred only transiently (early or late) during the interval¹⁶. However, our findings are consistent with those of a study of working memory for faces also reported in this issue¹⁷.

One possible concern might be that the processes within dorsolateral PFC were transient, but lasted somewhat longer than sensory or motor-related processes. This might have produced a more prolonged haemodynamic response that did not have time to recover within the 10-s period between stimuli (see Methods), and thus would appear as a sustained response. To address this concern, we conducted a small supplemental study using the same

task, but with a 20-s delay to allow for recovery of the haemodynamic response. We observed sustained elevation of the functional magnetic resonance imaging (fMRI) signal within dorsolateral PFC even with this longer delay, consistent with our interpretation of processing that endures over the delay within this brain region.

The effect of load in PFC appears as a step function, with activation increasing primarily between the 1- and 2-back conditions (see legend to Fig. 1). Similar observations have been made in a positron emission tomography (PET) study using this task¹⁸. However, in at least one other study we have observed a more gradual (linear) increase of PFC activation across levels of load¹¹, and in the current study a similar graded pattern was observed within the areas of parietal cortex showing an effect of load. At present, we do not fully understand these discrepancies, although the step pattern observed within PFC may be related to the fact that the 2- and 3-back conditions depend on maintenance of information about the

Table 1 Regions exhibiting significant task-related activity

Region (Brodmann's area)	Talairach coordinates (x, y, z)	Statistical effect (Max z-value)
Prefrontal cortex		
R middle frontal gyrus (BA9/46)	37 32 30	Load (3.82)
R superior frontal gyrus (BA8)	7 29 37	Load (3.52)
L, R inferior frontal gyrus (BA44)	-41 8 16, 45 13 26	Load (4.26, 3.15)
L inferior frontal gyrus (BA44)	-43 8 26	Load × time (4.91)
Cingulate cortex		
Anterior cingulate (BA32)	0 5 40 0 16 43	Time (5.61) Load × time (2.93)
Insular cortex		
L, R frontal operculum	-29 17 1, 32 22 4	Time (4.02, 3.72)
Motor and somatosensory cortex		
L, R middle frontal gyrus (BA6)	-35 2 40, 32 6 58	Load (3.01, 4.12)
R middle frontal gyrus (BA6)	46 5 45	Load × time (2.87)
L, R middle frontal gyrus (BA6)	-26 -4 56, 31 -1 54	Load × time (3.09, 3.03)
L precentral gyrus (BA6)	-45 -4 42	Load × time (2.45)
L precentral gyrus (BA4)	-35 -15 55	Time (5.20)
L postcentral gyrus (BA1/2/3)	-45 -25 57	Time (4.75)
Parietal cortex		
R superior parietal lobule (BA7)	12 -61 53	Load (3.37)
L, R supramarginal gyrus (BA40)	-40 -48 40, 44 -55 43	Load (3.72, 3.75)
L, R supramarginal gyrus (BA40)	-28 -41 41, 37 -44 39	Load × time (3.26, 4.34)
Temporal cortex		
R superior temporal gyrus (BA22)	64, -28 14	Time (3/41)
Visual cortex		
L precuneus (BA18)	-20 -66 29	Time (3.34)
L, R cuneus/lingual gyrus (BA17/18)	-9 63 4, 7 -60 4	Time (3.96, 4.05)
L middle occipital gyrus (BA19)	-37 -60 -3	Time (3.21)

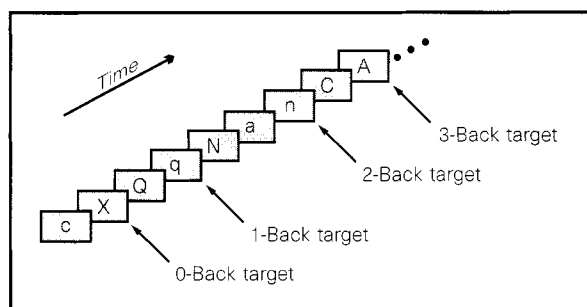


Figure 1 The four memory conditions of the sequential letter ('n-back') task. In the 0-back conditions, subjects responded to a single pre-specified target (such as X). In the '1-back' condition, the target was any letter identical to the one immediately preceding it (one trial back). In the 2-back and 3-back conditions, the target was any letter identical to the one presented 2 or 3 trials back, respectively. Thus, working memory load increased from the 0-back to the 3-back conditions. All other task parameters were the same across all conditions.

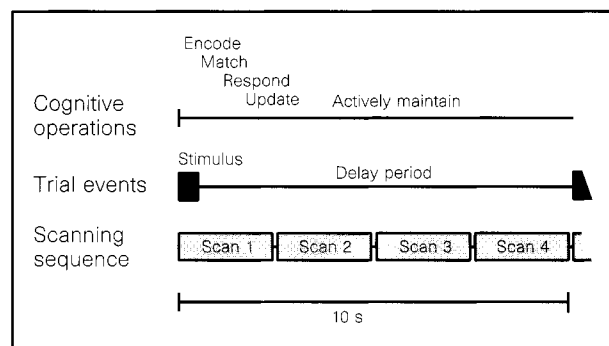


Figure 2 Diagram of the experimental protocol, showing the sequence of stimulus events and scan acquisition during a single trial, as well as the approximate time course of mental operations presumed to be engaged by the task.

sequential order of stimuli, whereas the 0- and 1-back conditions do not.

The finding of sustained activation within dorsolateral PFC challenges the strong view that there is a clean dissociation between executive and maintenance processes, with PFC housing the former but not the latter. At the very least it suggests that the relationship between executive functions and active maintenance may be more complex than originally thought. For example, recent computational theories^{2,19} suggest that PFC may contribute to executive control by actively maintaining particular types of information (such as goal or other context representations that bias processing in posterior systems, or the sequential relationships between stimuli), or that it may be required for active maintenance under particular circumstances (such as in the presence of distracting information²⁰). These possibilities are also consistent with recent neurophysiological findings of sustained activity of neurons within the PFC of non-human primates in delayed response tasks^{5,6}, some of which have specifically involved the maintenance of sequential order information^{7,21} and distractor stimuli⁸. Thus, a characterization of the type of information that is actively maintained or the conditions

under which maintenance is required may be more relevant to the function of PFC than a distinction between active maintenance and executive control.

Alternatively, PFC may not be a site of active storage but may still play an adjunct role by issuing periodic control signals that continue to engage more posterior systems in which information is actively maintained. If the control signal were generated sufficiently often (and more so at higher levels of load), it might produce a pattern of sustained, load-sensitive activation within PFC as measured by fMRI. Thus, differing views regarding the role that PFC plays in active maintenance may be consistent with the pattern of dorsolateral PFC activation observed in our study. However, these views make different claims about whether or not task-specific information is actually maintained within PFC. An adjudication of this issue will require more detailed data than has been provided by human neuroimaging studies to date.

A different possibility is that sustained activation within PFC reflects the operation of more general processes associated with task difficulty and mental effort, rather than processes specific to working memory. We examined this possibility in another recently

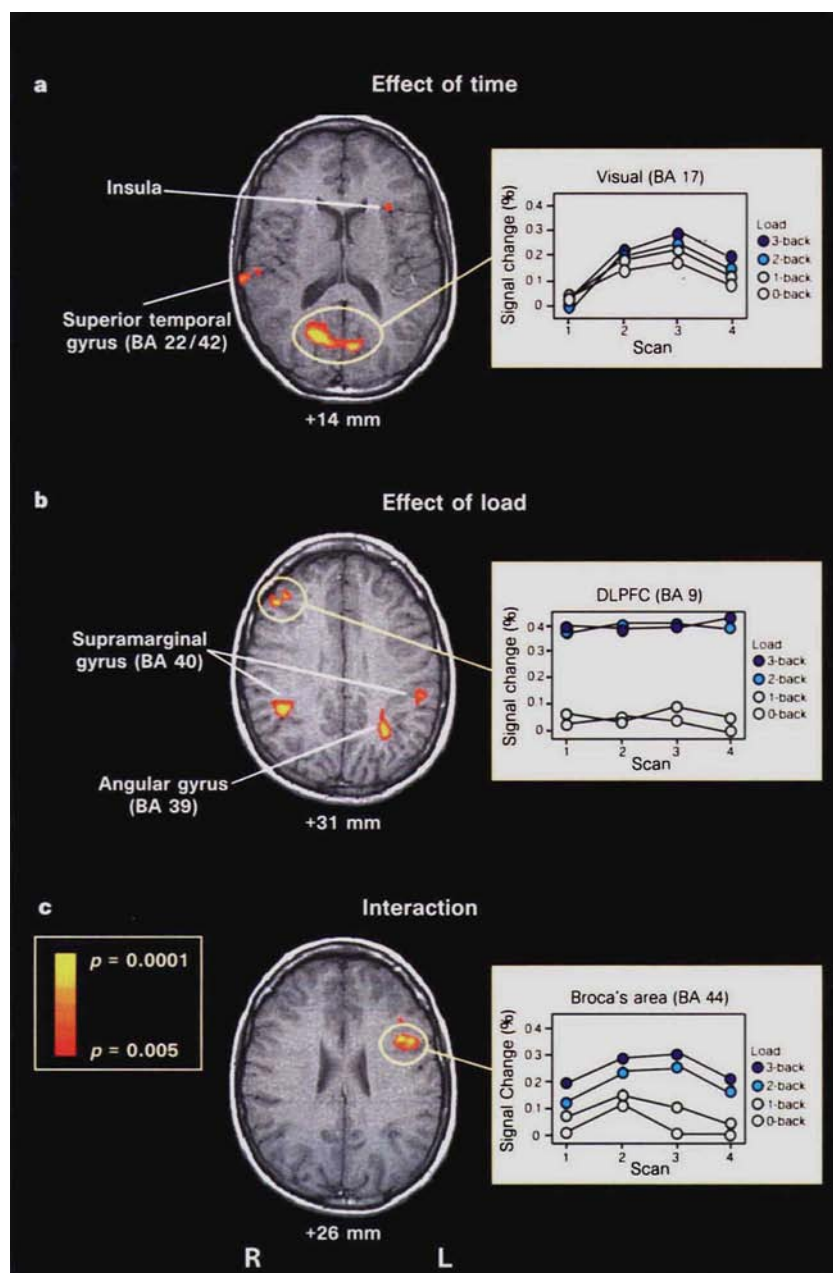


Figure 3 Images showing regions of activation representative of each effect. Insets plot the signal for the pixel showing the strongest effect in the designated region, as a proportion of the lowest value of that pixel across scans and load conditions. **a**, Effect of time; **b**, effect of load; **c**, Interaction between time and load.

completed experiment²², in which memory and task difficulty were independently manipulated. Dorsolateral PFC was activated in memory-demanding conditions (showing a similar sustained pattern), but not in other conditions that relied significantly less on active maintenance but were more difficult (for example, stimulus degradation). Thus, activation of this structure seems to reflect processes specific to working memory, and not processes associated more generally with task difficulty.

Apart from the dorsolateral PFC, sensitivity to load was observed primarily in posterior regions of frontal cortex (including Broca's area), and in posterior parietal cortex. These areas may participate with dorsolateral PFC in the active maintenance of information within working memory^{10,23}. However, unlike dorsolateral PFC, some of these structures exhibited additional regions of activation characterized by an interaction of load and time. All of the latter showed transient activation that was greater and more prolonged as load increased, similar to the pattern shown in the inset for Fig. 3c. These systems may play a role in transient working memory processes, such as the updating of contents, comparison operations and possibly rehearsal. The last of these is of particular interest with regard to the pattern of activation observed in Broca's area (left BA44; see Fig. 3c), considering the role that this region is thought to play in articulatory rehearsal^{13,4,23,24}. Rehearsal is usually assumed to be an ongoing process, suggesting that we should have observed a pattern of continuous activation in Broca's area. However, introspective reports from the subjects suggest that they may have stopped rehearsal early in the delay in the lower-load conditions of our experiment. This would be consistent with the observation of transient activation at lower loads and more prolonged activation at higher loads. This interpretation requires further examination. Nevertheless, the pattern of activation observed within Broca's area is distinct from the one observed in dorsolateral PFC, and raises the possibility of a dissociation between the processes underlying explicit verbal rehearsal, and other mechanisms for actively maintaining information that may reside within dorsolateral PFC. □

Methods

Cognitive task. Stimuli were pseudorandom sequences of 14 consonants, presented visually (500 ms duration, 9,740 ms interstimulus interval) using a Macintosh computer and PsyScope software²⁵. The 10 s trial duration was chosen to allow sufficient time for the MR signal to return to baseline between trials for stimulus-locked events^{12,13} and to permit acquisition of fMRI data over four time periods during each trial (see Fig. 2). Subjects used their dominant hand to respond to each stimulus, by pressing one button for targets (33% of trials) and another for non-targets. Only subjects who demonstrated acceptable accuracy (>75%) on all conditions during a pretesting session were scanned. During scanning, average accuracy across conditions was 90% for target trials and 98% for non-target trials, which is comparable to previous studies using trials of shorter duration. The task was administered in blocks of 14 trials at constant load level, during which scanning occurred (see Fig. 2). Six blocks were run for each of the four levels of load, pseudo randomly ordered (with the constraint that each level was sampled once every four blocks), to control for confounding effects of time on task, head movement and scanner drift.

Image acquisition. Scanning took place in a conventional 1.5T GE Signa whole-body and standard RF coil scanner in the UPMC MR Research Center. Twenty-four slices (3.75 mm³ isotropic voxels) were acquired parallel to the AC-PC line. Functional scans, T2* weighted for maximum sensitivity to variations in oxygen-saturation^{14,26} were acquired using a 4-interleave spiral-scan pulse sequence²⁷ (TR 640 ms, TE 35 ms, flip angle 40°, FOV 24 cm). This allowed 8 slices to be acquired every 2.5 s. Scanning was synchronized with stimulus presentation so that a set of 8 slices was acquired four times during each 10 s trial. The same set of 8 slices was scanned for three consecutive trials, followed by a different set of 8 slice locations. Set order was counterbalanced across blocks to control for asynchronous acquisitions across regions. Scanning occurred during only 9 of the 14 trials in each block, with the remainder used to load working memory and allow the fMRI signal to achieve steady state. Nine trials per block × 6 blocks per condition divided by 3 scan sets per whole brain

yielded a total of 18 scans in each slice plane (1–24) for each time point (1–4) at each level of load (0–3). Following functional scanning, a high-resolution structural scan was performed in the same planes as the functional scan for anatomic localization and coregistration of images across subjects (discussed below).

Image analysis. Before analysis, images for all subjects were coregistered to a common reference brain, using a 12-parameter automated algorithm (AIR²⁸), and then smoothed using a three-dimensional gaussian filter (8 mm FWHM) to accommodate between-subject differences in anatomy. This approach to pooling data across subjects is used frequently in PET studies to increase statistical power and permit quantitative identification of regions that activate reliably across subjects. We have used this approach successfully in a previous fMRI study¹¹. Planned contrasts were used to identify pixels showing any one of the following three effects ($P < 0.005$): (1) monotonic increase in signal as a function of load; (2) transient increase in signal over time, greater during scans 2 and 3 of each trial than scans 1 and 4; or (3) a combination of these effects that manifested as a significant interaction of time and load (determined using a 2-way ANOVA followed by the contrasts described for both of the first two effects). The second of these contrasts (used to identify transient increases associated with stimulus and response-locked events) takes into account the expected 5 s lag in the fMRI response, based on observations that we and others have made regarding the time course of the fMRI signal in response to brief events^{12,13}. Images were formed for each of the three effects that included pixels significant only for the corresponding effect. Regions made up of 8 or more contiguous pixels were then identified in each of these images (as a precaution against type I errors²⁹, and insuring an effective pixel-wise alpha of $P < 0.005$). Finally, these regions were overlaid onto the reference structural images (to determine their anatomic location), which were then transformed to standard stereotactic space using AFNI software³⁰.

Received 12 September 1996; accepted 7 February 1997.

1. Baddeley, A. D. *Working Memory* (Oxford Univ. Press, New York, 1986).
2. Shallice, T. *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, Cambridge, 1988).
3. Gathercole, S. E. Neuropsychology and working memory: a review. *Neuropsychology* **8**, 494–505 (1994).
4. Paulesu, E., Frith, C. D. & Frackowiak, R. S. J. The neural correlates of the verbal component of working memory. *Nature* **362**, 342–345 (1993).
5. Fuster, J. M. & Alexander, G. E. Neuron activity related to short-term memory. *Science* **173**, 652–654 (1971).
6. Goldman-Rakic, P. S. in *Handbook of Physiology: The Nervous System* (eds Plum, F. & Mountcastle, V.) 373–417 (American Physiological Society, Bethesda, Maryland, 1987).
7. Barone, P. & Joseph, J. P. Prefrontal cortex and spatial sequencing in macaque monkey. *Exp. Brain Res.* **78**, 447–464 (1989).
8. Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
9. Cohen, J. D. et al. Activation of prefrontal cortex in a nonspatial working memory task with functional MRI. *Hum. Brain Mapp.* **1**, 293–304 (1994).
10. Smith, E. E., Jonides, J. & Koeppe, R. A. Dissociating verbal and spatial working memory using PET. *Cereb. Cortex* **6**, 11–20 (1996).
11. Braver, T. S. et al. A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage* **6**, 49–62 (1997).
12. Savoy, R. L. et al. in *Society of Magnetic Resonance, 3rd Meeting* 450 (International Society of Magnetic Resonance in Medicine, Nice, 1995).
13. Vazquez, A. L. & Noll, D. C. Non-linear temporal aspects of the BOLD response in fMRI. *Proc. Int. Soc. Magn. Res. Med. 4th Meet.* 1765 (1996).
14. Kwong, K. K. et al. Dynamic magnetic resonance of human brain activity during primary sensory stimulation. *Proc. Nat. Acad. Sci. USA* **89**, 5675–5679 (1992).
15. Tulving, E. et al. Neuroanatomical correlates of retrieval in episodic memory: auditory sentence recognition. *Proc. Nat. Acad. Sci. USA* **91**, 2012–2015 (1994).
16. Fiez, J. A. et al. A positron emission tomography study of the short-term maintenance of verbal information. *J. Neurosci.* **16**, 808–822 (1996).
17. Courtney, S. M., Ungerleider, L. G. & Keil, K. & Haxby, J. V. Transient and sustained activity in a distributed neural system for human working memory. *Nature* (this issue).
18. Jonides, J. et al. Verbal working memory load affects regional brain activity as measured by PET. *J. Cogn. Neuroscience* (in the press).
19. Cohen, J. D., Braver, T. S. & O'Reilly, R. A computational approach to prefrontal cortex, cognitive control, and schizophrenia: recent developments and current challenges. *Phil. Trans. R. Soc. Lond. B* **351**, 1515–1527 (1996).
20. D'Esposito, M. et al. The neural basis of the central executive system of working memory. *Nature* **378**, 279–281 (1995).
21. Petrides, M. E. Functional specialization within the dorsolateral frontal cortex for serial order memory. *Proc. R. Soc. Lond. B* **246**, 293–298 (1991).
22. Barch, D. M. et al. Dissociating working memory from task difficulty in human prefrontal cortex. *Neuropsychologia* (in the press).
23. Awh, E. et al. Dissociation of storage and rehearsal in verbal working memory: evidence from PET. *Psychol. Sci.* **7**, 25–31 (1996).
24. Waters, G., Rochon, E. & Caplan, D. The role of high-level speech planning in rehearsal: evidence from patients with apraxia of speech. *J. Mem. Lang.* **31**, 54–73 (1992).
25. Cohen, J. D., MacWhinney, B., Flatt, M. R. & Provost, J. PsyScope: a new graphic interactive environment for designing psychology experiments. *Behav. Res. Meth. Instrum. Comput.* **25**, 257–271 (1993).
26. Ogawa, S. et al. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping using MRI. *Proc. Nat. Acad. Sci. USA* **89**, 5951–5955 (1992).
27. Noll, D. C., Cohen, J. D., Meyer, C. H. & Schneider, W. Spiral K-space MR imaging of cortical activation. *J. Magn. Reson. Imag.* **5**, 49–56 (1995).

28. Woods, R. P., Cherry, S. R. & Mazziotta, J. C. Rapid automated algorithm for aligning and reslicing PET images. *J. Comp. Assist. Tomogr.* **16**, 620–633 (1992).
29. Forman, S. D. *et al.* Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magn. Reson. Med.* **33**, 636–647 (1995).
30. Cox, R. W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162–173 (1996).

Acknowledgements. We thank D. Barch and B. J. Casey for their thoughtful comments and suggestions, and F. Sabb for his invaluable assistance in the conduct of this study. This work was supported by grants from the US National Institute of Mental Health, National Institute on Aging, National Institute on Neurological Disorders and Stroke, Office of Naval Research and the McDonnell-Pew Program in Cognitive Neuroscience.

Correspondence and requests for materials should be addressed to J.D.C. (e-mail: jdcohen@cmu.edu).

Transient and sustained activity in a distributed neural system for human working memory

Susan M. Courtney*, Leslie G. Ungerleider†, Katrina Keil* & James V. Haxby*

* Section on Functional Brain Imaging, † Section on Neurocircuitry, Laboratory of Brain and Cognition, 10 Center Drive, MSC 1366, National Institutes of Health, Bethesda, Maryland 20892-1366, USA

Working memory involves the short-term maintenance of an active representation of information so that it is available for further processing. Visual working memory tasks, in which subjects retain the memory of a stimulus over brief delays, require both the perceptual encoding of the stimulus and the subsequent maintenance of its representation after the stimulus is removed from view. Such tasks activate multiple areas in visual and prefrontal cortices^{1–9}. To delineate the roles these areas play in perception and working memory maintenance, we used functional magnetic resonance imaging (fMRI) to obtain dynamic measures of neural activity related to different components of a face working memory task—non-selective transient responses to visual stimuli, selective transient responses to faces, and sustained responses over memory delays. Three occipitotemporal areas in the ventral object vision pathway had mostly transient responses to stimuli, indicating their predominant role in perceptual processing, whereas three prefrontal areas demonstrated sustained activity over memory delays, indicating their predominant role in working memory. This distinction, however, was not absolute. Additionally, the visual areas demonstrated different degrees of selectivity, and the prefrontal areas demonstrated different strengths of sustained activity, revealing a continuum of functional specialization, from occipital through multiple prefrontal areas, regarding each area's relative contribution to perceptual and mnemonic processing.

Imaging studies in humans have shown that both visual and prefrontal cortices are active during visual working memory tasks, but these studies did not have sufficient temporal resolution to distinguish transient activity during stimulus presentation from sustained activity after the stimulus is removed from view^{1–9}. Increases in activity in visual and prefrontal cortices could therefore be attributed either to attentional enhancement of responses to stimuli, or to working memory, or to both. By contrast, electrophysiological studies in humans, which do have sufficient temporal resolution to distinguish transient from sustained activity, have not had sufficient spatial resolution to attribute the sources of these signals to specific anatomical locations^{10,11}. We used fMRI to measure changes in neural activity over the course of a working memory task in order to distinguish the perceptual and mnemonic roles played by the cortical regions that participate in visual working memory (see also ref. 12).

Functional MRI scans were obtained while eight subjects alternately performed working memory and sensorimotor control tasks (Fig. 1; and see Methods). Six bilateral regions were identified that were consistently activated across subjects (Tables 1 and 2). The temporal responses of these regions were analysed using multiple regression with regressors related to three different cognitive components of the task: (1) a transient, non-selective response to visual stimulation that was equivalent for faces and scrambled faces; (2) a transient, selective response to faces; and (3) a sustained response over memory delays (Figs 1–3; and see Methods). Because complex objects other than faces were not used as stimuli, the transient, selective response to faces could indicate either face-selectivity *per se* or selectivity for any meaningful visual stimulus. Similarly, a sustained response over memory delays could indicate either a working memory system that is specific to faces, or a generic visual object working memory system.

All subjects showed activity in ventral occipitotemporal extrastriate visual areas that was correlated with stimulus presentation. A region in the posterior lingual and fusiform gyri, Brodmann area (BA) 18, showed a transient, mostly non-selective response to stimuli (Fig. 2a). In half of the subjects, the response to faces in this region was significantly greater than the response to scrambled faces, but the magnitude of this difference was small. Rostral to this relatively non-selective area, a ventral temporal region in the mid-to-anterior fusiform gyrus (BA 37) showed a transient, more face-selective response with a small, but significant, level of sustained activity during the memory delay. This pattern of activation indicates that the anterior fusiform gyrus is involved to a greater extent in the perceptual processing of faces than is the more posterior region and may also be involved in the maintenance of an active representation of the face during the memory delay (Fig. 2b). Face-selective responses, similar to the responses in the anterior fusiform gyrus, were also seen in the inferior occipital sulcus (BA 18/19), an area more lateral and more posterior than the fusiform face-selective area. The activity in the inferior occipital sulcus was more sustained during the memory delay than was the activity in the anterior fusiform gyrus, but the regression weights for non-selective and face-selective responses to stimuli in these regions were approximately the same.

Three distinct prefrontal areas were identified that all showed sustained activity during the memory delay interval: one in the posterior middle and inferior frontal gyri (BA 9/44), a second in the inferior frontal gyrus near the anterior end of the insula (BA 45/47), and a third in the anterior middle frontal gyrus (BA 46) (Table 2; Figs 2c–e and 3). The contribution of the memory delay regressor was significant for all three of these areas, but the relative magnitudes of the regression weights were different, indicating that the functional responses for these regions differed. The posterior middle frontal gyrus area had the most non-selective visual stimulation activity and the least memory delay activity, whereas the anterior middle frontal area had the least non-selective visual stimulation activity and the most memory delay activity. Regression weights for the posterior middle frontal gyrus differed significantly from those for the other two prefrontal areas ($F(1, 111) = 27.91$, $P < 0.0001$). Although the difference between the weights for the inferior frontal and anterior middle frontal areas did not reach significance ($P = 0.054$), their anatomical segregation clearly identified them as distinct regions.

There was a systematic progression in the relative strengths of perception- and memory-related activity from posterior extrastriate through prefrontal areas, indicating that this distributed neural system for working memory is hierarchically organized (Fig. 3). In the extrastriate visual areas, the progression from mainly non-selective perceptual to face-selective perceptual activity is consistent with the well established hierarchical organization of visual cortex¹³. Progressive changes in perception- and memory-related activity in the prefrontal areas suggest that these areas, like those in extrastriate

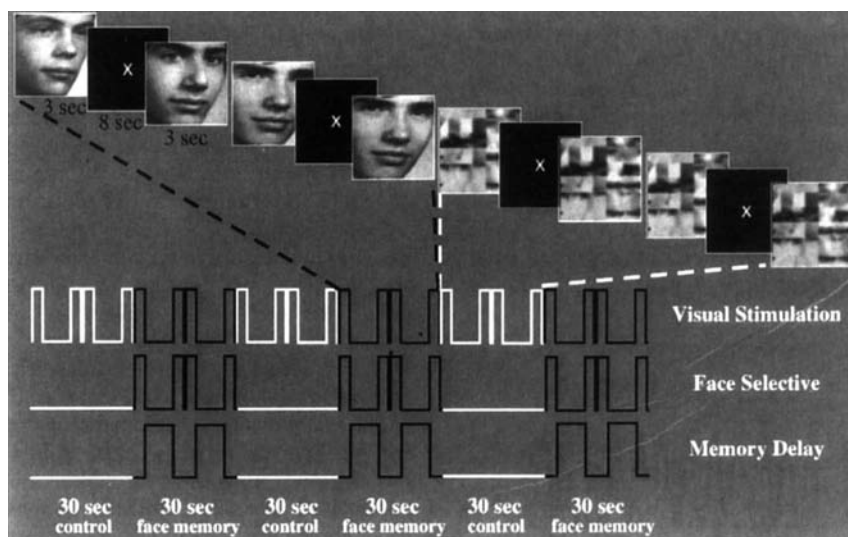


Figure 1 For each series of scans, subjects performed three baseline-activation task cycles, each consisting of 30 s of a sensorimotor control task, followed by 30 s of a working-memory task. Each task period consisted of two items for that task. Three time series are shown which represent the different cognitive components of the task: a transient, non-selective response to visual stimuli; a

transient, selective response to faces; and sustained activity during memory delays. These time series (smoothed and delayed by convolution with a model of the haemodynamic response) were used as regressors in a multiple regression analysis of the time course of activation in each area (see Methods).

Table 1 Areas showing statistically significant activation in at least half of the subjects

Area of activation	Subjects							
	S1	S2	S3	S4	S5	S6	S7	S8
Posterior lingual and fusiform gyri (BA 18)	B	B	B	B	B	B	B	B
Mid-to-anterior fusiform gyrus (BA 37)	B	B	L	R	B	B	B	-
Inferior occipital sulcus (BA 18/19)	B	B	B	-	B	B	-	B
Posterior mid-frontal/inferior-frontal gyrus (BA9/44)	L	R	-	-	R	B	L*	L*
Anterior insula/inferior frontal gyrus (BA 45/47)	B	B	B	-	L	B	-	-
Anterior mid-frontal gyrus (BA 46)	L	B	-	-	B	B	L*	-

Numbers in parentheses are the probable Brodmann areas

L, left hemisphere only; R, right hemisphere only; B, bilateral.

* Voxels in these regions reached the threshold of $Z > 3.09$, but the region of activation was not of sufficient spatial extent to pass the correction for multiple comparisons.

Table 2 Group analysis of areas of activation

Area of activation	Talairach coordinates			Multiple regression coefficients			R ²
	x	y	z	stimul†	faces†	delay†	
Posterior lingual and fusiform gyri (BA 18)	19 − 24	− 90 − 90	− 11 − 6	0.78*** 0.72***	0.17* 0.21*	0.05 0.07	0.763 0.771
Mid-to-anterior fusiform gyrus (BA 37)	31 − 32	− 60 − 60	− 20 − 16	0.32* 0.40**	0.52*** 0.40**	0.16* 0.20**	0.587 0.682
Inferior occipital sulcus (BA 18/19)	34 − 28	− 85 − 82	1 − 1	0.27** 0.50*	0.45** 0.28*	0.28*** 0.22*	0.583 0.442
Posterior mid-frontal/inferior frontal gyrus (BA 9/44)	42 − 38	13 15	26 25	0.19 0.15*	0.55*** 0.37*	0.26* 0.47***	0.550 0.489
Anterior insula/inferior frontal gyrus (BA 45/47)	30 − 31	21 20	0 3	0.17* 0.04	0.33*** 0.46*	0.51*** 0.50***	0.496 0.465
Anterior mid-frontal gyrus (BA 46)	28 − 20	43 40	16 11	0.11 − 0.17	0.34*** 0.50***	0.55*** 0.67***	0.503 0.505

Numbers in parentheses are the probable Brodmann areas (BA)³⁰. A positive x Talairach coordinate³⁰ indicates an area in the right hemisphere; a negative x coordinate indicates an area in the left hemisphere. There were no significant hemispheric effects ($P > 0.1$ for main effect and all interactions). Asterisks, contribution of regressor is significant in ≥ 50 , ≥ 75 or 100% of active regions for one, two or three asterisks, respectively.

† Each regression coefficient is divided by the sum of all three coefficients.

R^2 , fraction of variance accounted for by all 3 regressors.

visual cortex, may also be hierarchically organized for working memory, although corroborating evidence about interregional connectivity and functional dependencies does not yet exist.

The precise roles played by the prefrontal regions in working memory remain to be specified. Previous imaging studies have reported activation in regions with similar locations during verbal and visual working memory tasks¹⁻⁹, as well as during long-term memory retrieval^{14,15}. Activation of the same areas by both long-term retrieval and working memory is consistent with the idea that retrieval produces an active representation of the recalled material much like the active representation of material held during a

working memory delay. Functional differences among several regions in prefrontal cortex have been studied^{7,8}. Our inferior prefrontal area (BA 45/47) is close to a ventrolateral prefrontal area previously identified⁷ which has been proposed to be involved in the encoding and retrieval of information held in posterior cortical areas. Verbal semantic functions have also been associated with this inferior frontal region^{16,17}, suggesting that its activation in our study may also reflect verbal processing of the face stimuli. Our anterior midfrontal area (BA 46) appears to be inferior and anterior to a cluster of regions in BA 9/46 that Petrides and colleagues^{7,8} have associated with monitoring mnemonic performance. Our posterior middle and inferior prefrontal region is inferior and anterior both to a region in BA 8 reported by Petrides⁸ and to the human frontal eye field¹⁸, suggesting that our result is related to neither conditional associative memory nor oculomotor control.

The reliability of results from this new method of fMRI time series analysis is corroborated by results from a subsequent face working memory study using different subjects and a different experimental protocol¹⁹. In the new study, subjects were required to remember a set of three faces instead of a single face, and a longer intertrial interval (6 s instead of 1 s) followed the test stimulus. This new study showed activation in the same three prefrontal areas that were identified in the current study and confirmed their relative participation in perceptual and mnemonic processing. The new study

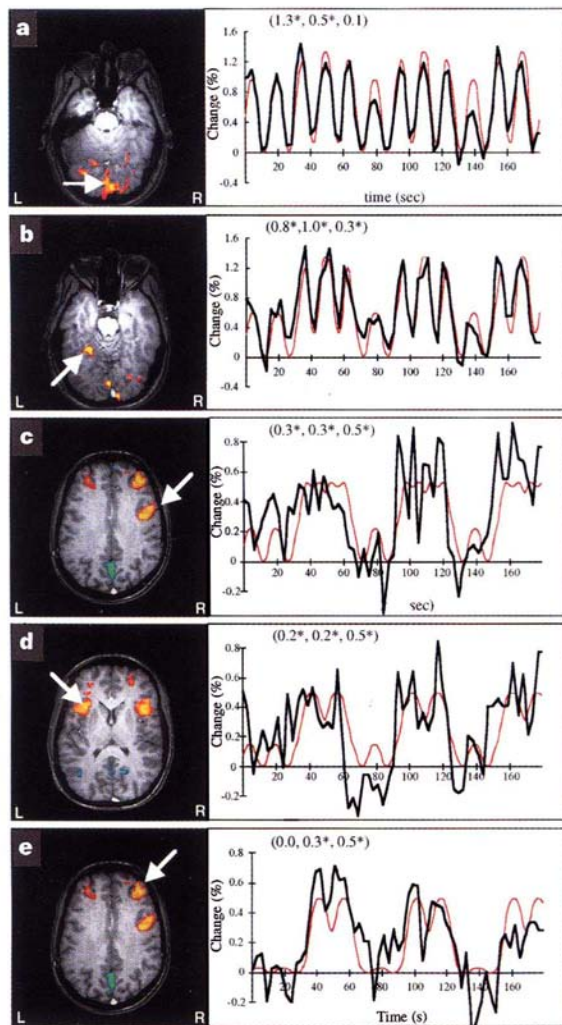


Figure 2 Results from a single subject, S2, overlaid onto that subject's anatomical MR images, collected in the same session and in the same planes as the functional MR images. Red and yellow indicate areas with significant positive correlations with either visual stimulation, memory task, or both. Areas shown in blue and green have significant negative correlations. Per cent changes in signal over time, averaged over all voxels in the activated region, are shown by the thick black line to the right of each MRI image. Time series were analysed with the multiple regression model described for Fig. 1. The thin red line shows the sum of the regressors, multiplied by their corresponding coefficients, shown in parentheses above each graph (non-selective perceptual, face-selective perceptual, memory delay). Asterisks indicate regressors whose contribution was statistically significant ($P < 0.05$). **a**, Posterior lingual and fusiform gyri (BA 18). **b**, Mid-to-anterior fusiform gyrus (BA 37). **c**, Posterior mid- and inferior frontal gyri (BA 9/44). **d**, Inferior frontal gyrus and anterior insula (BA 45/47). **e**, Anterior mid-frontal gyrus (BA 46).

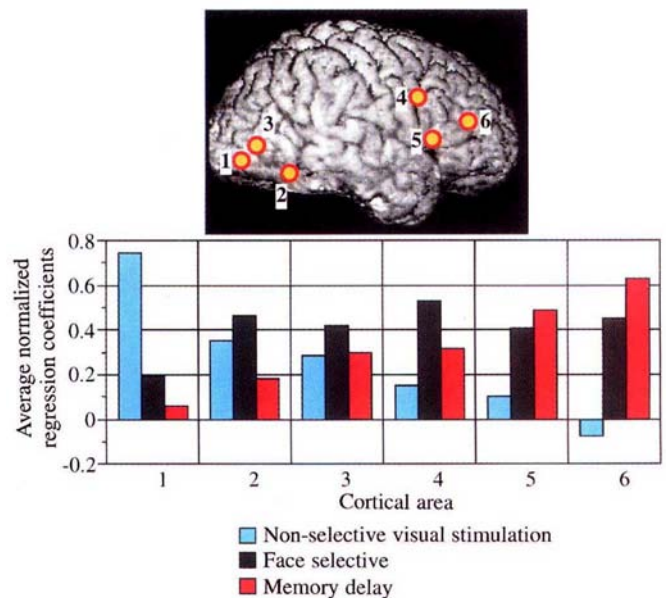


Figure 3 Average locations across subjects of the regions of activation listed in Tables 1 and 2 are shown in lateral projection onto a brain that has been stereotactically normalized into Talairach space²⁰. Locations shown are the mean locations for the right and left hemispheres. The areas, numbered in order of decreasing relative contribution from non-selective visual stimulation and increasing contribution from the memory delay, are: (1) posterior lingual and fusiform gyri (BA 18), (2) mid-to-anterior fusiform gyrus (BA 37), (3) inferior occipital sulcus (BA 18/19), (4) posterior mid- and inferior frontal gyri (BA 9/44), (5) inferior frontal gyrus and anterior insula (BA 45/47), and (6) anterior mid-frontal gyrus (BA 46). The graph shows the three normalized regression coefficients for each area, averaged across subjects and across hemispheres, demonstrating the gradual shift in the relative contributions of each task component to the activity within this distributed neural system.

also confirmed that the sustained activity during the memory delay was related to working memory and not simply to a more prolonged response to faces than to control stimuli. Because the new study had a much longer intertrial interval following the test stimulus, prolonged activity during this interval, related solely to the perceptual processing of faces, could be distinguished from sustained activity during the working memory delay interval. Activity during the working memory delay was significantly higher than activity following the test face in the posterior middle frontal area (increases of 0.42 versus 0.03% relative to control item intervals, $P < 0.001$) and the anterior middle frontal area (increases of 0.57 versus 0.09%, $P < 0.001$). The effect in the inferior frontal area was in the expected direction but failed to reach statistical significance (increases of 0.34 versus 0.21%, $P > 0.1$).

These fMRI results demonstrate the existence of areas in human cortex with response properties that are analogous to those described in single-cell recording experiments in non-human primates: (1) early extrastriate visual areas demonstrate transient, relatively non-selective responses to complex visual stimuli²⁰; (2) later extrastriate visual areas demonstrate transient, selective responses to faces, indicating a more specialized role in the processing of meaningful images^{21,22}; and (3) both extrastriate visual^{23,24} and prefrontal^{25,26} cortical areas demonstrate sustained activity during memory delays, indicating a role in maintaining an active representation of the face in working memory. In non-human primates, delay activity in temporal cortex, unlike that in prefrontal cortex, is disrupted by intervening stimuli²⁷. These results, supported by our results in humans, suggest that the dominant function of temporal cortex is perceptual but that this region also participates in maintaining a working memory as long as it is not recruited for the perception of new stimuli. Our results extend findings from non-human primates both by quantifying the relative strengths of sustained activity during delays in later extrastriate visual as compared to prefrontal areas and by demonstrating that in human cortex, multiple, functionally distinct, prefrontal regions participate in working memory. □

Methods

Task. Each memory item began with a 3-s presentation of a single sample face, followed by an 8-s fixation delay, and then a 3-s presentation of a test face. Subjects were instructed to keep an image of the sample face in mind during the delay, and to respond by pressing a right button to indicate that the test face matched the immediately preceding face and a left button if it did not. For the control task, which controlled for visual stimulation, motor response, and anticipation of response, subjects saw a scrambled face, a fixation delay, and then the same scrambled face, after which they responded by pressing both buttons. The control task used the same timing as the memory task. Subjects did not attempt to remember or match the scrambled faces. Scrambled faces were filtered to remove the high-frequency edges created by scrambling.

Imaging. 12 contiguous, axial, 6-mm-thick slices were obtained in 8 series of 64 scans each (repeat time, 3 s), beginning at the most ventral extent of the occipital lobe in 8 healthy volunteers (3 male, 5 female; age 24 ± 2.1 yr, education 16.6 ± 1.1 yr). Gradient echo, echo planar imaging was used (TE 40 ms, flip angle 90°) on a GE Signa 1.5 Tesla magnet. A significant increase in the MRI signal is referred to in the text as 'activation'. All subjects gave written informed consent.

Statistics. Voxels that were activated during the working memory task were identified by calculating correlations between the time series of MRI intensities in a single voxel and two idealized response functions²⁸, reflecting contrasts between (1) non-selective visual stimulation versus no stimulation, and (2) the memory task versus the control task (face selective + memory-delay regressors; Fig. 1). To increase statistical power, all 8 series of scans in each subject were analysed together using an ANCOVA to factor out the series variance. All statistical results have a single voxel Z threshold of 3.09 (degrees of freedom corrected for correlation between adjacent time points). Statistical significance ($P < 0.05$) of a region of activation was determined using an analysis based on the spatial extent of each region to correct for multiple comparisons²⁹. The time

course of activation in each of these regions was analysed using multiple regression²⁸. Each regressor was tested to see whether it could account for a significant portion of the variance, independent of the variance attributable to the other regressors. A weighting coefficient for each regressor was obtained by simultaneously finding the best fit to the data for the sum of the three regressors multiplied by their weighting coefficients. The weights produced by multiple regression provide measures of the relative strengths of (1) perceptually non-selective, (2) perceptually selective, and (3) memory-related responses.

Because of variability in the magnitude of responses, both across subjects and across regions, regression coefficients were normalized by dividing each coefficient by the sum of all three. These normalized regression coefficients express the relative strength of three types of responses in a single region. Regional differences in normalized regression weights were tested by treating each significantly activated region as an independent observation in a 3-way ANOVA (region $\times$ hemisphere $\times$ regressor).

Received 4 September 1996; accepted 11 February 1997.

- Cohen, J. D. *et al.* Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI. *Hum. Brain Mapp.* **1**, 293–304 (1994).
- Courtney, S. M., Ungerleider, L. G., Keil, K. & Haxby, J. V. Object and spatial working memory activate separate neural systems in human cortex. *Cereb. Cortex* **6**, 39–49 (1996).
- D'Esposito, M. *et al.* The neural basis of the central executive system of working memory. *Nature* **378**, 279–281 (1995).
- Fiez, J. A. *et al.* A positron emission tomography study of the short-term maintenance of verbal information. *J. Neurosci.* **16**, 808–822 (1996).
- Haxby, J. V., Ungerleider, L. G., Horwitz, B., Rapoport, S. I. & Grady, C. L. Hemispheric differences in neural systems for face working memory: A PET rCBF study. *Hum. Brain Mapp.* **3**, 68–82 (1995).
- McCarthy, G. *et al.* Functional magnetic resonance imaging of human prefrontal cortex activation during a spatial working memory task. *Proc. Natl Acad. Sci. USA* **91**, 8690–8694 (1994).
- Owen, A. M., Evans, A. C. & Petrides, M. Evidence for a two-stage model of spatial working memory processing within the lateral frontal cortex: a positron emission tomography study. *Cereb. Cortex* **6**, 31–38 (1996).
- Petrides, M., Alivisatos, B., Evans, A. C. & Meyer, E. Dissociation of human mid-dorsolateral from posterior dorsolateral frontal cortex in memory processing. *Proc. Natl Acad. Sci. USA* **90**, 873–877 (1993).
- Smith, E. E., Jonides, J. & Koepp, R. A. Dissociating verbal and spatial working memory using PET. *Cereb. Cortex* **6**, 11–20 (1996).
- Chao, L. L. & Knight, R. T. Prefrontal and posterior cortical activation during auditory working memory. *Cog. Brain Res.* **4**, 27–37 (1996).
- Gevins, A. *et al.* High resolution evoked potential imaging of the cortical dynamics of human working memory. *Electroenceph. Clin. Neurophys.* **98**, 327–348 (1996).
- Cohen, J. D. *et al.* Temporal dynamics of brain activation during a working memory task. *Nature* **386**, 604–608 (1997).
- Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
- Buckner, R. L. Beyond HERA: Contributions of specific prefrontal brain areas to long-term memory retrieval. *Psychonom. Bull. Rev.* **3**, 149–158 (1996).
- Haxby, J. V. *et al.* Face encoding and recognition in the human brain. *Proc. Natl Acad. Sci. USA* **93**, 922–927 (1996).
- Demb, J. B. *et al.* Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *J. Neurosci.* **15**, 5870–5878 (1995).
- McCarthy, G., Blamire, A. M., Rothman, D. L., Gruetter, R. & Shulman, R. G. Echo-planar magnetic resonance imaging studies of frontal cortex activation during word generation in humans. *Proc. Natl Acad. Sci. USA* **90**, 4952–4956 (1993).
- Paus, T. Location and function of the human frontal eye-field: a selective review. *Neuropsychologia* **34**, 475–483 (1996).
- Courtney, S. M., Maisog, J. M., Ungerleider, L. G. & Haxby, J. V. Extrastriate and frontal contributions to face and location working memory. *Soc. Neurosci. Abstr.* **22**, 968 (1996).
- Desimone, R. & Ungerleider, L. G. in *Handbook of Neuropsychology* (eds Boller, F. & Grafman, J.) Vol. 2, 267–300 (Elsevier, Amsterdam, 1989).
- Perrett, D. I., Rolls, E. T. & Caan, W. Visual neurons responsive to faces in the monkey temporal cortex. *Exp. Brain Res.* **47**, 329–342 (1982).
- Desimone, R., Albright, T. D., Gross, C. G. & Bruce, C. Stimulus selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* **4**, 2051–2062 (1984).
- Miyashita, Y. & Chang, H. S. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature* **331**, 68–70 (1988).
- Miller, E. K., Li, L. & Desimone, R. Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neurosci.* **13**, 1460–1478 (1993).
- Fuster, J. M. *Memory in the Cerebral Cortex: An Empirical Approach to Neural Networks in the Human and Nonhuman Primate* (MIT Press, Cambridge, Massachusetts, 1995).
- Wilson, F. A., O'Scalaidhe, S. P. & Goldman-Rakic, P. S. Dissociation of object and spatial processing domains in primate prefrontal cortex. *Science* **260**, 1955–1957 (1993).
- Miller, E. K., Erickson, C. A. & Desimone, R. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *J. Neurosci.* **16**, 5154–5167 (1996).
- Friston, K. J. *et al.* Analysis of fMRI time-series revisited. *NeuroImage* **2**, 45–53 (1995).
- Friston, K. J., Worsley, K. J., Frackowiak, R. S. J., Mazziotta, J. C. & Evans, A. C. Assessing the significance of focal activations using their spatial extent. *Hum. Brain Mapp.* **1**, 210–220 (1994).
- Talairach, J. & Tournoux, P. *Co-Planar Stereotaxic Atlas of the Human Brain* (transl. Rayport, M.) (Thieme Medical, New York, 1988).

Acknowledgements. We thank A. Martin and R. Desimone for comments on an earlier draft of the manuscript; V. Clark, F. Lalonde and P. Jezzard for assistance with fMRI data collection; A. Karni and J. Maisog for physician coverage; and J. Maisog for assistance with computer programming and data analysis.

Correspondence and requests for materials should be addressed to S.M.C. (e-mail: susan_courtney@nih.gov).

Allosteric activation and tuning of ligand efficacy in cyclic-nucleotide-gated channels

Gareth R. Tibbs*, Evan H. Goulding* & Steven A. Siegelbaum*†

* Center for Neurobiology and Behavior, † Department of Pharmacology, Howard Hughes Medical Institute, Columbia University, 722 W. 168 Street, New York, New York 10032, USA

Despite recent advances in the identification of ligand-binding^{1,2} and voltage-sensing³ regions of ion channels, the domains that couple such regions to channel opening have not been identified. Moreover, it is uncertain whether ligand binding or depolarization are obligatory steps that must precede channel opening (according to linear reaction schemes^{4,5}) or whether they act to stabilize the channel in an open state that can exist independently of ligand binding or depolarization (according to cyclic allosteric models^{6–8}). By comparing ligand-independent and ligand-dependent channel openings, we now show that retinal and olfactory cyclic-nucleotide-gated channels² are activated by a cyclic allosteric mechanism. We further show that an amino-terminal domain, distinct from the pore and ligand-binding motifs, participates in the allosteric gating transition, accounting for differences in the free energy of gating of the two channels. The allosteric transition provides an important mechanism for tuning the physiological response of ligand-binding proteins, such as cyclic-nucleotide-gated channels, to different biological signals.

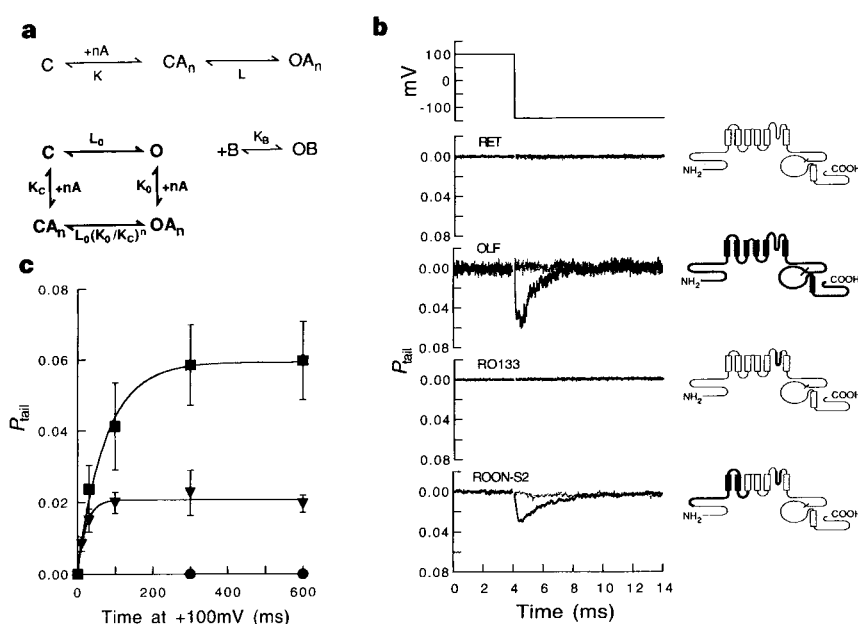
Retinal cyclic-nucleotide-gated (CNG) channels are selectively activated by cyclic GMP, exhibiting a higher sensitivity (lower $K_{1/2}$) and a higher maximum open probability ($P_{\max}$) to cGMP than to cAMP, reflecting the primary role of cGMP in visual signal transduction². In contrast, olfactory CNG channels are activated equally well by cAMP and cGMP, reflecting the importance of both

messengers in olfaction². By constructing chimaeras between retinal and olfactory channels, two domains responsible for their differences in ligand gating have been identified^{9–11}. Cyclic nucleotide binding takes place within a 120-amino-acid region in the carboxy terminus^{2,12} that is homologous to cyclic-nucleotide-binding (CNB) domains of other proteins¹³. The efficacy with which ligand binding opens the channels ($P_{\max}$) also depends on an amino terminus (N-S2) domain⁹. Thus, constructs that contain the catfish olfactory¹⁴ (OLF) N-S2 region exhibit a higher $P_{\max}$ than constructs that contain the same CNB domain but include the bovine retinal¹² (RET) N-S2 region. To explain these observations, we proposed that the N-S2 domain helps determine the free energy difference between the closed and open states of the channel (ΔG_{c-o}) and that the OLF N-S2 domain endows channels with a more favourable ΔG_{c-o} , increasing their tendency to open⁹.

Here we measure ligand-independent CNG channel openings to explore the role of the N-S2 domain in coupling ligand binding to channel activation and to distinguish between linear and cyclic models of gating (Fig. 1a). These experiments allow us to test the following three predictions of cyclic allosteric models: (1) both RET and OLF should show ligand-independent opening; (2) the ligand-independent spontaneous open probability (P_{sp}) of OLF should be significantly greater than that of RET (owing to the postulated lower ΔG_{c-o} of OLF); (3) the P_{sp} of a chimaeric channel should be determined by the identity of the N-S2 domain.

To enhance detection of ligand-independent openings, we exploited the voltage-dependent, open-channel block of CNG channels¹⁵ by the N-terminal inactivation 'ball' peptide¹⁶ of Shaker BK^+ channels (KK peptide¹⁷). At +100 mV, CNG channels that open spontaneously should accumulate in an open-KK peptide-blocked state (Fig. 1a), resulting in a tail current upon hyperpolarization as the channels rapidly unblock and transit through the open state. Tail currents, whose amplitudes increase exponentially with the time that the patch is held depolarized, are indeed observed in inside-out patches containing OLF channels (Fig. 1b, c). The lack of a tail current with RET (Fig. 1b) is consistent with the prediction of a lower P_{sp} than OLF, but could also reflect the lower affinity of RET for KK peptide¹⁵. To distinguish between these possibilities, we

Figure 1 Tail currents reveal ligand-independent channel openings. **a**, Models of channel gating and block. Top, linear activation model⁶. C, Closed channel; O, open channel; A, agonist molecule; K, ligand dissociation constant; L, gating equilibrium constant; n, number of binding events. Bottom, cyclic allosteric model⁶ (bold face). K_O and K_C are open and closed state ligand dissociation constants. L_0 is the allosteric equilibrium constant. Agonists activate the channel by binding more tightly to the open than the closed state. Binding of KK peptide (B), an open-channel blocker, to the unliganded open state (O) is shown in light face. K_B , voltage-dependent dissociation constant of KK peptide. **b**, Tail currents recorded in the presence (dark traces) and absence (light traces) of 200 μ M KK peptide. Top trace, voltage stepped to -140 mV following 300-ms step to +100 mV. Holding potential was -80 mV. Other panels, tail currents from patches containing ~2,400 RET, 180 OLF, 1,950 RO133 and 1,800 ROON-S2 channels, expressed as approximate open probability (P_{tail} ; see Methods). A 150- μ s period following the voltage step is blanked to eliminate a residual capacitance artefact. **c**, Peak value of P_{tail} as a function of time at +100 mV for OLF (■), ROON-S2 (▼) and RO133 (●) ($n = 22, 19$ and 18 patches, respectively; not all points determined in each patch). Data were fitted by single exponential functions (τ : OLF, 76 ms; ROON-S2, 28 ms).



studied a chimaeric RET channel, RO133, which has the P-region and pore properties (including KK peptide affinity) of OLF but retains the gating properties of RET^{15,18} (see Methods). Patches containing RO133 channels do not exhibit tail currents either (Fig. 1b), indicating that the P_{sp} values of both RO133 and RET are significantly lower than that of OLF.

To provide an independent and quantitative measure of P_{sp} , we next directly recorded spontaneous single-channel currents. Representative currents from an inside-out patch containing OLF chan-

nels are shown in Figs 2 and 3a. Several criteria indicate that such activity represents unliganded channel openings. First, the noise at 0 mV is well fitted by a single gaussian function, consistent with the reversal potential of these channels¹⁸. Second, the increased noise at other voltages is stable and absent from patches lacking cyclic-nucleotide-activated current (not shown). Third, the spontaneous openings display a characteristic voltage-dependent block by KK peptide (Figs 2, 3b). For OLF, the spontaneous channel currents yield a value for P_{sp} of 2.25×10^{-3} . This provides a value for the

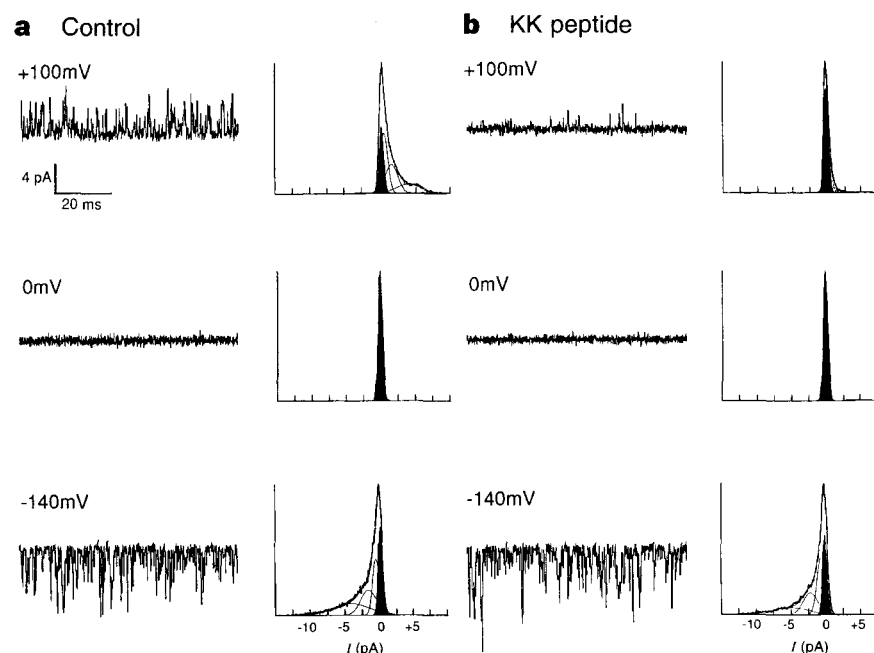


Figure 2 Ligand-independent openings of OLF. Sweeps from a single patch containing 180 OLF channels at -140 , 0 and $+100$ mV in the absence (a) and presence (b) of $200 \mu\text{M}$ KK peptide. All points amplitude histograms (constructed from 750-ms sweeps and fitted with multiple gaussians, smooth lines within the data) are displayed to the right of the appropriate current records (closed component is the filled gaussian).

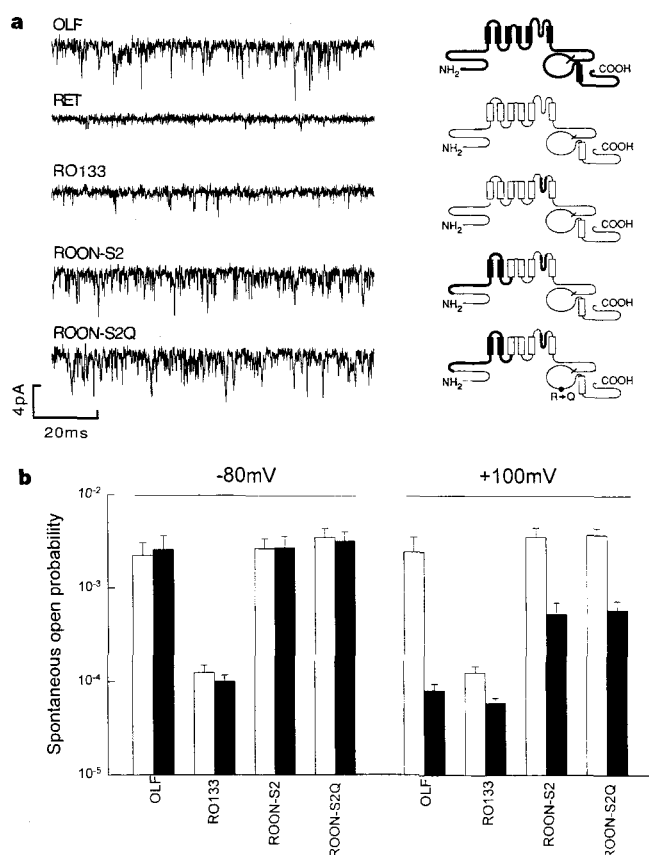


Figure 3 Spontaneous open probabilities determined from unliganded single channel activity. **a**, Representative spontaneous channel currents at ~ 80 mV for patches containing ~ 180 OLF, 500 RET, 660 RO133, 200 ROON-S2 and 105 ROON-S2Q channels. **b**, Logarithmically scaled histogram of mean spontaneous open probabilities (P_{sp}). White and black bars, P_{sp} in absence or presence of $200 \mu\text{M}$ KK peptide, respectively. Mean P_{sp} at -80 mV: OLF, $2.25 \times 10^{-3} \pm 0.84 \times 10^{-3}$, $n = 16$; ROON-S2, $2.69 \times 10^{-3} \pm 0.73 \times 10^{-3}$, $n = 9$; RO133, $1.25 \times 10^{-4} \pm 0.24 \times 10^{-4}$, $n = 23$; ROON-S2Q, $3.37 \times 10^{-3} \pm 0.91 \times 10^{-3}$, $n = 4$. OLF, ROON-S2 and ROON-S2Q were significantly different from RO133 ($P < 0.05$, ANOVA), but not from each other. Note: the somewhat lesser open-channel block of ROON-S2 (and ROON-S2Q) relative to OLF can explain the smaller tail current seen with ROON-S2 relative to OLF (Fig. 1b, c).

allosteric equilibrium constant, L_0 , between the unliganded closed (C) and open (O) states of the channel of 443, corresponding to a $\Delta G_{C \rightarrow O}$ of $3.6 \text{ kcal mol}^{-1}$.

Whereas RET or RO133 tail currents were not detectable, spontaneous openings are evident for both channels (Fig. 3a, and see also ref. 8). Because the low single-channel conductance of RET prevents a reliable estimation of its P_{sp} , we determined the P_{sp} of RO133, which is equal to 1.25×10^{-4} , 18-fold lower than the value for OLF (Fig. 3b). These data yield a $\Delta G_{C \rightarrow O}$ of $5.3 \text{ kcal mol}^{-1}$ for RO133, which is $1.7 \text{ kcal mol}^{-1}$ larger than that of OLF, confirming the second prediction of our model.

To determine whether the N-S2 region accounts for the difference in $\Delta G_{C \rightarrow O}$ between RET and OLF, we replaced the N-S2 domain of RO133 with that of OLF, generating the double chimera ROON-S2. When assayed with KK peptide, ROON-S2, like OLF, exhibits large tail currents whose amplitudes increase exponentially as a function of time at $+100 \text{ mV}$ (Fig. 1b, c). Measurements of spontaneous single-channel openings of ROON-S2 reveal a P_{sp} of 2.69×10^{-3} ($\Delta G_{C \rightarrow O} = 3.5 \text{ kcal mol}^{-1}$) nearly identical to that of OLF (Fig. 3). Although the poor expression of the reverse chimera ORN-S2 (RET N-S2 domain in OLF) precludes a quantitative estimate of its P_{sp} , its P_{max} (determined with 30 mM cGMP) is very small, approximately equal to the P_{sp} of OLF, and no unliganded openings are observed (data not shown). This suggests that the P_{sp} for ORN-S2 must be far lower than that of OLF, consistent with a RET-like value of L_0 .

To eliminate the possibility that the spontaneous openings are due to cyclic nucleotide contamination, we constructed a binding-site mutant, ROON-S2Q, in which arginine 536 was replaced by glutamine. R536 is homologous to an arginine residue that forms an important ionic interaction with the cyclized phosphate group of cAMP in homologous CNB domains^{19,20}. We find that this mutation reduces the apparent affinity (increases $K_{1/2}$) for cGMP 26-fold without altering P_{sp} relative to that of ROON-S2, arguing that the spontaneous openings are independent of agonist binding (Fig. 3).

We next asked how the differences in L_0 imposed by the RET and OLF N-S2 domains are expected to alter ligand-dependent gating and whether such changes are consistent with the observed differ-

ences in gating between RET and its chimaeras containing the OLF N-S2 domain. We used the Monod–Wyman–Changeux (MWC) concerted allosteric model⁶ to evaluate the effects of changes in L_0 (Fig. 1a) because this simple scheme requires only four parameters that can be either directly determined (L_0) or obtained from fits to dose–response curves (Fig. 4). A selective decrease in L_0 of RO133 to that of OLF is predicted to decrease the $K_{1/2}$ for cGMP (Fig. 4a) and increase P_{max} with cAMP (Fig. 4b), qualitatively similar to the observed changes in gating between RO133 and ROON-S2. However, the predictions underestimate the observed shifts in $K_{1/2}$ and P_{max} , suggesting that changing the N-S2 domain also alters the ligand-binding properties of the channel. This may indicate that distinct domains in the primary sequence (N-S2 and the CNB domain) contribute directly to ligand binding, similar to the proposed clam-shell-like binding site of the ionotropic glutamate receptors²¹. However, the similarity between CNB domains of the CNG channels and other CNB proteins whose structures are known¹³ suggests that the effect of the N-S2 domain is not direct, but is a consequence of its allosteric coupling to the binding site (Fig. 5), as observed in other allosteric proteins²².

Our results show that a single domain, the N-S2 region, determines differences in both ligand gating and spontaneous openings

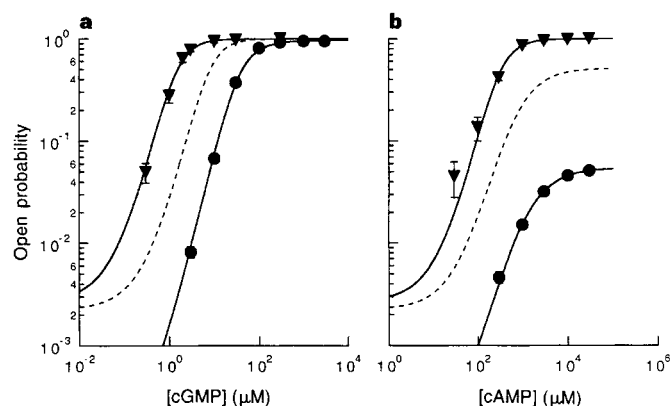


Figure 4 Observed and predicted cGMP and cAMP dose–response curves. Data points represent observed open probabilities as a function of ligand concentration plotted on a logarithmic scale for ROON-S2 ($\blacktriangledown$) and RO133 ($\bullet$). Solid lines represent fits of the MWC model to these data, assuming two binding sites as this gave a better fit to the data at low open probability compared to fits with four binding events (not shown). Dashed lines are the predicted dose–response relationships for the chimera ROON-S2 assuming that K_O and K_C are equal to those of RO133 (see below) and L_0 is equal to that of OLF (443; see text). **a**, Activation by cGMP. K_O , K_C and P_{max} from fits were $0.39 \mu\text{M}$, $170 \mu\text{M}$ and 0.960 for RO133 and $0.08 \mu\text{M}$, $22 \mu\text{M}$ and 0.995 for ROON-S2. **b**, Activation by cAMP. K_O , K_C and P_{max} from fits were $47 \mu\text{M}$, $1,015 \mu\text{M}$ and 0.054 for RO133 and $17 \mu\text{M}$, $3,475 \mu\text{M}$ and 0.990 for ROON-S2.

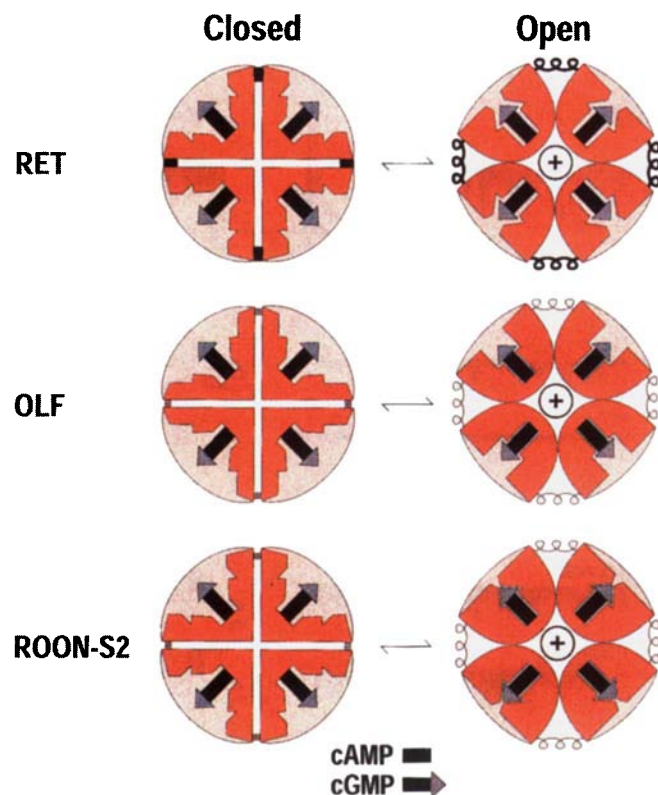


Figure 5 The coupling between $\Delta G_{C \rightarrow O}$ and ligand-binding affinity. Channel opening is allosterically coupled to a change in the binding site, which increases ligand affinity. The RET binding domain interacts strongly with cGMP and weakly with cAMP. The OLF binding domain accommodates cAMP and cGMP equally and hence interacts weakly with either ligand. The N-S2 domains are depicted as springs that mediate subunit–subunit interactions which stabilize the channel in the closed state. The coupling of OLF subunits by weak springs (compared to the strong springs between RET subunits) depicts the lower $\Delta G_{C \rightarrow O}$ imparted by the OLF N-S2 domain. Transfer of the weaker OLF springs to the RET channel in ROON-S2 permits the RET binding domain to adopt a more complementary conformation to cGMP and cAMP when the channel is open compared to its interaction with cGMP and cAMP in RET, accounting for enhanced ligand affinity of the N-S2 chimera (Fig. 4). Although not shown, changes in $\Delta G_{C \rightarrow O}$ can also influence closed-state binding.

between RET and OLF channels. This indicates that the spontaneous openings are part of the normal ligand-gating reaction (as distinct from a local perturbation of the pore²³) and that channel activation occurs by a cyclic allosteric mechanism. As regions homologous to the N-S2 domain are important for subunit assembly in voltage-gated K⁺ channels^{24–27}, a change in quaternary structure at subunit interfaces may underlie activation of voltage-gated channel family members, as has been proposed for ligand-gated channels²⁸. Furthermore, this analysis suggests how changes in the allosteric transition can tune ligand efficacy without altering the intrinsic structure of the binding site (Fig. 5). For example, the requirement that OLF must accommodate both cGMP and cAMP with equal affinity is likely to decrease the ability of its binding pocket to attain a high degree of complementarity with either ligand, limiting the binding energy available for gating. OLF channels compensate for this deficiency by having a relatively low ΔG_{c-o} . Second messenger systems such as Ca²⁺-calmodulin²⁹ and phosphorylation³⁰, may alter the profile of agonists that are capable of efficiently activating a channel by dynamically altering the value of the allosteric equilibrium. □

Methods

Molecular biology. Chimaeras and point mutations were synthesized by PCR and subcloning then verified using dideoxy chain termination^{9,18}. RO133: OLF S314–F348 replaced RET A344–A378. RON-S2: OLF E89–R215 replaced RET N91–S240. ROON-S2: RON-S2 in which OLF S314–F348 additionally replaced RET A344–A378. ORN-S2: RET V89–Y265 replaced OLF L87–V239. ROON-S2Q: ROON-S2 in which arginine 536 (R559 in RET) has been replaced by glutamine. In the diagrams, OLF sequences are bold, RET are light. Single channel conductance of P region chimaeras was identical to that of the P region donor¹⁸ (data not shown). Changing the P region did not alter ligand activation parameters. Thus, $K_{1/2}$ and P_{max} values for activation by cGMP were as follows, RET: $43 \pm 7 \mu M$ ($n = 3$); 0.913 ± 0.040 ($n = 5$), compared to RO133: $41 \pm 3 \mu M$ ($n = 7$); 0.948 ± 0.015 ($n = 6$). RON-S2: $3.0 \pm 0.4 \mu M$ ($n = 7$); 0.997 ± 0.003 ($n = 2$), compared to ROON-S2: $1.9 \pm 0.4 \mu M$ ($n = 9$); 0.999 ± 0.001 ($n = 5$). $K_{1/2}$ and P_{max} of OLF were $45 \pm 13 \mu M$ ($n = 6$) and 0.421 ± 0.128 ($n = 6$). $K_{1/2}$ and P_{max} of ROON-S2Q were $50 \pm 8 \mu M$ ($n = 8$) and 0.999 ± 0.001 ($n = 3$).

Electrophysiological recordings. Inside-out patches were obtained from *Xenopus* oocytes expressing homomeric channels composed of RET¹² or OLF¹⁴ α -subunits or their chimaeras 1–7 days after injection with cRNA (Message Machine, Ambion). Patches were excised into a high-flow-rate bath naive of cyclic nucleotide (symmetrical 67 mM KCl, 30 mM NaCl, 10 mM EGTA, 1 mM EDTA, 10 mM HEPES, pH 7.2). KK peptide (MAVAGLYGLGKKRQHRKKQ) is the Sh B inactivation peptide synthesized with two mutations¹⁷ to increase affinity for CNG channels¹⁵. For OLF (and constructs containing OLF P region), KK peptide dissociation constant is $\sim 0.01 \mu M$ at +100 mV and increasing e -fold for a 17–20 mV hyperpolarization¹⁵. Data were acquired using an Axopatch 200A patch clamp amplifier (Axon Instruments) with analogue leak and capacitance compensation. Tail currents were measured using a P/10 protocol to subtract uncompensated linear leak and capacitance. Two to eight identical tail current records were averaged per trace. Currents were low-pass-filtered at 30 kHz (eight-pole Bessel filter, Frequency Devices 902) and digitized at 200 kHz (Macintosh Centris 650 personal computer; ITC-16 interface and PULSE software, Instrutech). Open probability at the peak of the tail $P_{tail} = (I_{tail}/I_{max})P_{max}$, where I_{tail} is the uncorrected peak tail current at -140 mV; I_{max} is the maximal current at -140 mV (determined after completion of ligand-independent measurements by introduction of a second perfusion system into the bath to apply a saturating concentration of cGMP) and P_{max} was determined from single channel patches in the presence of saturating cGMP at -80 mV. As there is little voltage dependence to channel opening at saturating cGMP (conductance decreases e -fold for a 400 mV hyperpolarization and this is predominantly due to single channel current rectification; data not shown) P_{max} determined at -80 mV was used at other voltages without correction. The number of channels in a patch was determined from $I_{cGMP}/(iP_{max})$, all determined at -80 mV, where i is the single-channel current. Mean spontaneous currents (I_{sp}) due to unliganded single channel

openings (filtered at 4 kHz, digitized at 20 kHz) were determined from either the integral of the difference between the channel current and the zero current baseline (estimated by eye) or from the integral of all points amplitude distributions after subtracting the closed (baseline) gaussian component (see Fig. 2).

Data analysis. Activation parameters were determined as follows: $P_{sp} = (I_{sp}/I_{max})P_{max}$; $L_0 = [C]/[O] = (1/P_{sp}) - 1$; $\Delta G_{c-o} = -RT \ln[1/L_0]$. Observed open probabilities (P_{open}) were determined from I_{cGMP} (the current at a given cGMP concentration) from multichannel patches according to: $P_{open} = (I_{cGMP}/I_{max})(P_{max} - P_{sp}) + P_{sp}$, which corrects for subtraction of the spontaneous current from I_{cGMP} (all parameters determined at -80 mV). Values of the MWC model parameters K_O and K_C (μM) were derived from fits of the equation: $P_{open} = (1 + [A]/K_O)^n / (L_0(1 + [A]/K_O)^n + (1 + [A]/K_O)^n)$ to observed open probabilities, where L_0 , K_O , K_C and n are as defined in Fig. 1 legend. $K_{1/2}$ was determined by fits to the Hill equation $P_{open}/P_{max} = 1/(1 + (K_{1/2}/[A])^h)$, where h is the Hill coefficient. All data are presented as mean $\pm$ s.e.m.

Received 24 December 1996; accepted 10 February 1997.

- Karlin, A. & Akabas, M. H. Toward a structural basis for the function of nicotinic acetylcholine receptors and their cousins. *Neuron* **15**, 1231–1244 (1995).
- Zagotta, W. N. & Siegelbaum, S. A. Structure and function of cyclic nucleotide-gated channels. *Annu. Rev. Neurosci.* **19**, 235–263 (1996).
- Stuhmer, W. *et al.* Structural parts involved in activation and inactivation of the sodium channel. *Nature* **339**, 597–603 (1989).
- Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. (Lond.)* **117**, 500–544 (1952).
- del Castillo, J. & Katz, B. Interaction at end-plate receptors between different choline derivatives. *Proc. R. Soc. Lond. B* **146**, 369–381 (1957).
- Monod, J., Wyman, J. & Changeaux, J.-P. On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.* **12**, 88–118 (1965).
- Jackson, M. B. Spontaneous openings of the acetylcholine receptor channel. *Proc. Natl Acad. Sci. USA* **81**, 3901–3904 (1984).
- Picones, A. & Koren, J. I. Spontaneous, ligand-independent activity of the cGMP-gated ion channels in cone photoreceptors of fish. *J. Physiol. (Lond.)* **485**, 699–714 (1995).
- Goulding, E. H., Tibbs, G. R. & Siegelbaum, S. A. Molecular mechanism of cyclic-nucleotide-gated channel activation. *Nature* **372**, 369–374 (1994).
- Varnum, M. D., Black, K. D. & Zagotta, W. N. Molecular mechanism for ligand discrimination of cyclic nucleotide-gated channels. *Neuron* **15**, 619–625 (1995).
- Gordon, S. E. & Zagotta, W. N. Localization of regions affecting an allosteric transition in cyclic nucleotide-activated channels. *Neuron* **14**, 857–864 (1995).
- Kaupp, U. B. *et al.* Primary structure and functional expression from complementary DNA of the rod photoreceptor cyclic GMP-gated channel. *Nature* **342**, 762–766 (1989).
- Shabb, J. B. & Corbin, J. D. Cyclic nucleotide-binding domains in proteins having diverse functions. *J. Biol. Chem.* **267**, 5723–5726 (1992).
- Goulding, E. H. *et al.* Molecular cloning and single-channel properties of the cyclic nucleotide-gated channel from catfish olfactory neurons. *Neuron* **8**, 45–58 (1992).
- Kramer, R. H., Goulding, E. H. & Siegelbaum, S. A. Potassium channel inactivation peptide blocks cyclic nucleotide-gated channels by binding to the conserved pore domain. *Neuron* **12**, 655–662 (1994).
- Zagotta, W. N., Hoshi, T. & Aldrich, R. W. Restoration of inactivation in mutants of Shaker potassium channels by a peptide derived from ShB. *Science* **250**, 568–571 (1990).
- Murrell-Lagnado, R. D. & Aldrich, R. W. Energetics of Shaker K channels block by inactivation peptides. *J. Gen. Physiol.* **102**, 977–1003 (1993).
- Goulding, E. H., Tibbs, G. R., Liu, D. & Siegelbaum, S. A. Role of H5 domain in determining pore diameter and ion permeation through cyclic nucleotide-gated channels. *Nature* **364**, 61–64 (1993).
- Weber, I. T. & Steitz, T. A. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution. *J. Mol. Biol.* **198**, 311–326 (1987).
- Bubis, J., Neitzel, J. J., Saraswat, L. D. & Taylor, S. S. A point mutation abolishes binding of cAMP to site A in the regulatory subunit of cAMP-dependent protein kinase. *J. Biol. Chem.* **263**, 9668–9673 (1988).
- Stern-Bach, Y. *et al.* Agonist selectivity of glutamate receptors is specified by two domains structurally related to bacterial amino acid-binding proteins. *Neuron* **13**, 1345–1357 (1994).
- Perutz, M. *Mechanisms of Cooperativity and Allosteric Regulation in Proteins* 10–12 (Cambridge University Press, Cambridge, 1990).
- Auerbach, A., Sigurdson, W., Chen, J. & Akk, G. Voltage dependence of mouse acetylcholine receptor gating: Different charge movements in di-, mono- and unliganded receptors. *J. Physiol. (Lond.)* **494**, 155–170 (1996).
- Li, M., Jan, Y. N. & Jan, L. Y. Specification of subunit assembly by the hydrophilic amino-terminal domain of the Shaker potassium channel. *Science* **257**, 1225–1230 (1992).
- Shen, N. V., Chen, X., Boyer, M. M. & Pfaffinger, P. J. Deletion analysis of K⁺ channel assembly. *Neuron* **11**, 67–76 (1993).
- Babila, T., Moscucci, A., Wang, H., Weaver, F. E. & Koren, G. Assembly of mammalian voltage-gated potassium channels: evidence for an important role of the first transmembrane segment. *Neuron* **12**, 615–626 (1994).
- Tu, L. W. *et al.* Voltage-gated K⁺ channels contain multiple intersubunit association sites. *J. Biol. Chem.* **271**, 18904–18911 (1996).
- Unwin, N. Acetylcholine receptor channel imaged in the open state. *Nature* **373**, 37–43 (1995).
- Liu, M., Chen, T. Y., Ahmed, B., Li, J. & Yau, K. W. Calcium-calmodulin modulation of the olfactory cyclic nucleotide-gated cation channel. *Science* **266**, 1348–1354 (1994).
- Gordon, S. E., Brautigan, D. L. & Zimmerman, A. L. Protein phosphatases modulate the apparent agonist affinity of the light-regulated ion channel in retinal rods. *Neuron* **9**, 739–748 (1992).

Acknowledgements. We thank A. Karlin, R. Kass, D. Liu, P. Paoletti, L. Role, B. Santoro, A. Yamamoto and R. Taylor for helpful comments on the manuscript, C. Lam for help in preparing the figures, and J. Riley and H. Yao for technical assistance.

Correspondence and requests for materials should be addressed to G.R.T. (e-mail: grrt@columbia.edu).

Release of gelatinase A during platelet activation mediates aggregation

Grzegorz Sawicki*, Eduardo Salas*, Jesus Murat*, Helena Miszta-Lane* & Marek W. Radomski*†

Departments of * Pharmacology and † Obstetrics and Gynaecology, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Blood platelets limit blood loss at sites of vascular injury by forming a mechanical plug. They are also involved in thrombosis, atherosclerosis, inflammation and metastasis. Platelet activation is essential for these physiological and pathological reactions and depends upon their adhesion to the vessel wall and attachment to each other in the aggregation process. The two known pathways of aggregation are mediated by the release of endoperoxides/thromboxane A₂ and ADP¹⁻³ which amplify platelet aggregation. Here we report the identification of a new pathway of aggregation which is mediated by the release of a metalloproteinase enzyme, gelatinase A.

Gelatinase A (also known as matrix metalloproteinase-2, MMP-2 or type IV collagenase; EC 3.4.24.24) has a relative molecular mass of 72,000 (72K) and belongs to the group of matrix metalloproteinases that are involved in remodelling the extracellular matrix. Gelatinase A is secreted as a proenzyme (the 72K gelatinase) which is activated to a 64K form⁴⁻⁷. This secretion results from cell-matrix interactions and depends upon activation of adhesion receptors, including integrins^{8,9}. We now investigate whether human platelets express gelatinase A and if its release affects platelet aggregation.

Zymography and western blotting were used to identify metalloproteinases in platelets isolated from human blood. Zymography showed that small amounts of the 72K gelatinase were released by intact platelets (0.3 ± 0.1 ng enzyme protein per 10^8 platelets; mean $\pm$ s.d., $n = 8$) as calculated from the dose-response curve of recombinant human 72K gelatinase. This release was enhanced upon cell lysis (17.29 ± 3.72 ng enzyme; mean $\pm$ s.d., $n = 8$). The enzyme activity was inhibited by *o*-phenanthroline^{5,10} and SC44463 (ref. 11), both of which are selective inhibitors of metalloproteinases. Western blotting confirmed that the 72K gelatinase was present in the lysate of washed platelets (Fig. 1a, b). The 72K gelatinase, but not its activated form (Fig. 1a, c), was detected in the platelet releasate. This is, however, not surprising because gelatinase A is secreted as a pro-enzyme and is activated by a specific cell-surface event^{4-7,12}. Thus, it is likely that activation of platelet gelatinase A is linked to its translocation during aggregation to cell membrane fractions.

Collagen and thrombin caused concentration-dependent release of this enzyme to a maximum value of 10.59 ± 2.43 and 16.46 ± 2.93 ng per 10^8 platelets (mean $\pm$ s.d., $n = 5$), respectively (Fig. 2). This agrees with earlier work showing that the collagenolytic activity of human platelets¹³, as well as the release of 92K gelatinase from guinea-pig platelets¹⁴, is associated with platelet aggregation.

To investigate the significance of gelatinase A in platelet aggregation, monoclonal anti-gelatinase A antibodies, the recombinant human-tissue inhibitor of metalloproteinases-2 (TIMP-2), two pharmacologically selective inhibitors of metalloproteinases (phenanthroline and SC44463) and recombinant human gelatinase were used. The antibodies, TIMP-2, phenanthroline and SC44463 all inhibited collagen-induced aggregation (Fig. 3a-c) and release of ATP from platelets. There was a linear correlation between the inhibition of aggregation and the release of ATP induced both by

antibodies ($r = 0.99$, $P < 0.05$, $n = 5$), and phenanthroline ($r = 0.975$, $P < 0.05$, $n = 5$). Phenanthroline (100 μ M) was a weak inhibitor of platelet aggregation induced by phorbol myristate acetate (PMA) and fibrinogen, giving $11 \pm 7\%$ (mean $\pm$ s.d., $n = 4$) inhibition of aggregation. As PMA did not significantly increase the release of 72K gelatinase from intact platelets (0.6 ± 0.4 ng; $P > 0.05$, $n = 3$), these results show that metalloproteinase inhibitors affect aggregation through inhibition of gelatinase A. Phenanthroline and anti-gelatinase A antibodies both inhibited platelet aggregation in blood, their maximal effects being 42.8 ± 9.0 and $42.6 \pm 4.6\%$ (mean $\pm$ s.d., $n = 5$), respectively. The impedance assay of platelet aggregation in blood also detects accumulation of other blood elements on the electrodes¹⁵, and this may not be affected by inhibition of gelatinase A. The recombinant human activated, but not latent, gelatinase A, amplified aggregation triggered by collagen (Fig. 3d).

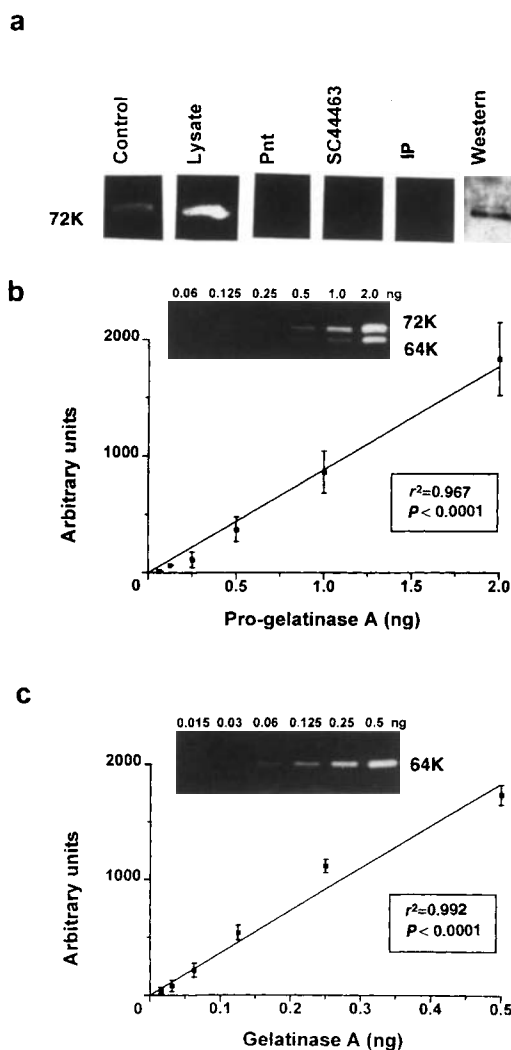


Figure 1 Detection and quantification of gelatinase A release from human platelets. **a**, The release of pro-gelatinase A was low in intact platelets (control). More enzyme was detected in the platelet lysate. The activity of pro-gelatinase A was reduced by phenanthroline (Pnt) and SC44463 and following removal of the enzyme by immunoprecipitation (IP). Pro-gelatinase A was also detected by western blotting. Results are representative of 4 separate experiments. **b**, **c**, Measurement of recombinant human pro-gelatinase A and activated gelatinase A. Zymography showed both pro-gelatinase A (72K) and its activated form (64K) in the preparation of recombinant human gelatinase A used in the experiments (**b**). The activated gelatinase A accounted for ~10% of total enzymatic protein as calculated from **c**. Following activation of pro-gelatinase A, only 64K gelatinase A was detected (**c**).

We also tested the effects of prostacyclin and nitric oxide (NO), which are inhibitors of aggregation, on the release of 72K gelatinase. A nitric oxide donor, S-nitroso-N-acetyl-DL-penicillamine (SNAP), and prostacyclin inhibited the secretion of gelatinase A induced by collagen and thrombin in intact platelets, indicating that enzyme release is controlled by NO and prostacyclin (Fig. 4a, b). There was a linear correlation between these inhibitor activities of SNAP and prostacyclin and their abilities to inhibit thrombin-induced aggregation ($r = 0.99$, $P < 0.05$, $n = 5$ for SNAP and $r = 0.98$, $P < 0.05$, $n = 5$ for prostacyclin).

In summary, 72K gelatinase is released during aggregation induced by collagen and thrombin, but not by PMA, and there is a positive correlation between its release and platelet aggregation; these effects are inhibited by prostacyclin and NO, and there is a positive correlation between inhibition of gelatinase A release and decreased aggregation; both monoclonal anti-gelatinase A antio-

dies, which neutralize gelatinase A activity, and the endogenous specific inhibitor of MMPs that preferentially inhibits gelatinase A, TIMP-2 (refs 16, 17), inhibit aggregation; selective inhibitors of metalloproteinase decrease aggregation at concentrations similar to those required for inhibition of gelatinase A; low concentrations of recombinant activated (but not latent) gelatinase A stimulate aggregation. These results show that the liberation of gelatinase A mediates platelet aggregation *in vitro* and suggest that TIMP-2 acts as an endogenous regulator of the pathway of platelet aggregation operated by gelatinase A.

The relationship between gelatinase A and the established pathways of aggregation was studied using aspirin and apyrase, which inhibit the production of thromboxane A₂ (ref. 18) and the action of ADP¹⁹ on platelets. Phenanthroline inhibited aspirin- and apyrase-insensitive aggregation (Fig. 4c), indicating that the pro-aggregatory effect of gelatinase A is independent of thromboxane and ADP

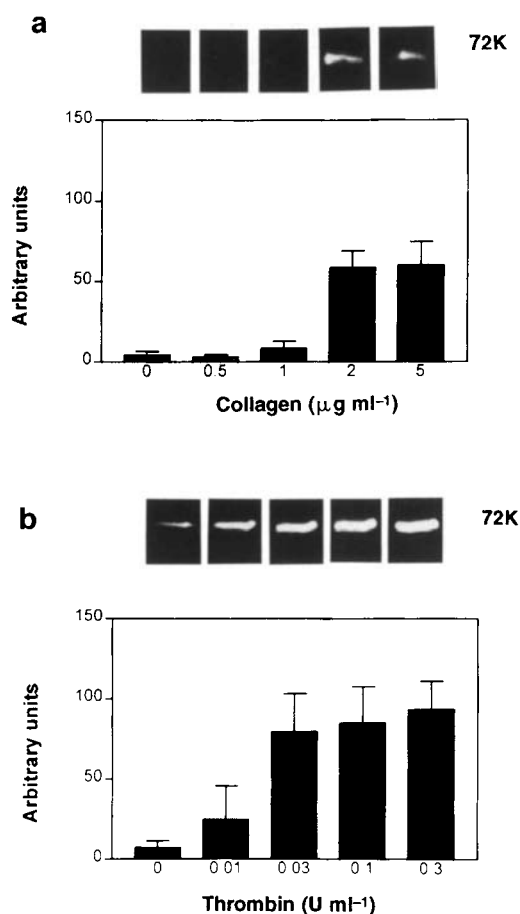


Figure 2 Release of gelatinase A induced by platelet aggregation. Both collagen (**a**) and thrombin (**b**) caused concentration-dependent release of pro-gelatinase A but not of gelatinase A ($< 15 \text{ pg}$). Less pro-gelatinase A was released by collagen than by thrombin, which released maximal amounts of enzyme. Bars are mean $\pm$ s.d. from 5–7 experiments and the corresponding gels show the data from a representative experiment.

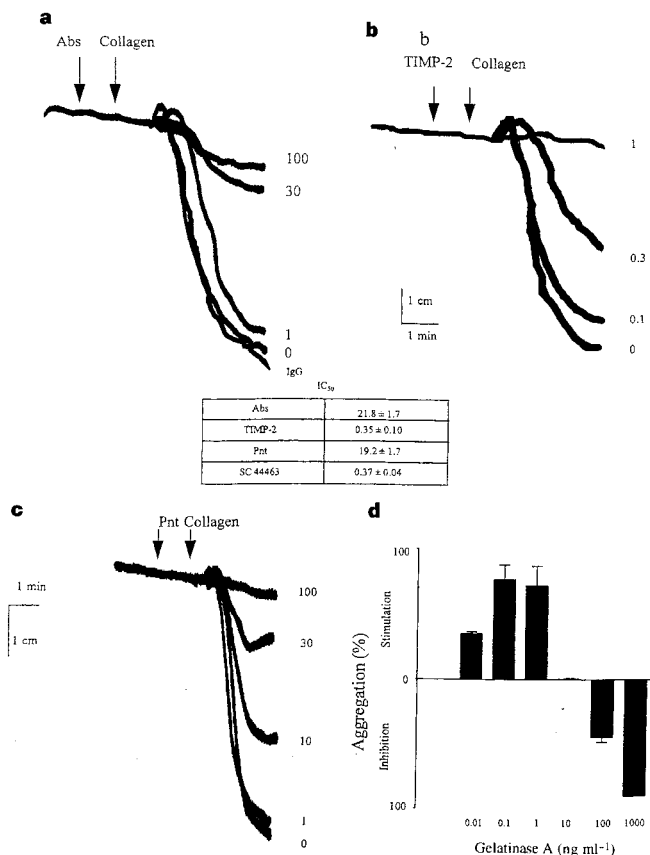


Figure 3 a–c, Aggregation of washed platelets. Numbers beside the traces show concentrations of anti-gelatinase A antibodies (Abs) and TIMP-2 in $\mu\text{g ml}^{-1}$, and of phenanthroline (Pnt), in μM . IgG: control aggregation in the presence of $100 \mu\text{g ml}^{-1}$ unrelated antibodies (mouse IgG₁; Calbiochem). Insert, inhibition of collagen-induced aggregation by monoclonal anti-gelatinase A antibodies, TIMP-2, phenanthroline and SC44463 (IC₅₀ values). **d**, Biphasic effect of recombinant gelatinase A on aggregation. Collagen ($0.5 \mu\text{g ml}^{-1}$)-induced aggregation was potentiated by low concentrations of gelatinase A but not by pro-gelatinase A (data not shown). In contrast, high concentrations of activated enzyme inhibited aggregation induced by collagen ($1\text{--}2 \mu\text{g ml}^{-1}$). Recombinant gelatinase A ($\leq 1,000 \text{ ng ml}^{-1}$; $n = 4$) did not cause aggregation of resting human platelets. Bars are mean $\pm$ s.d.; $n = 4$.

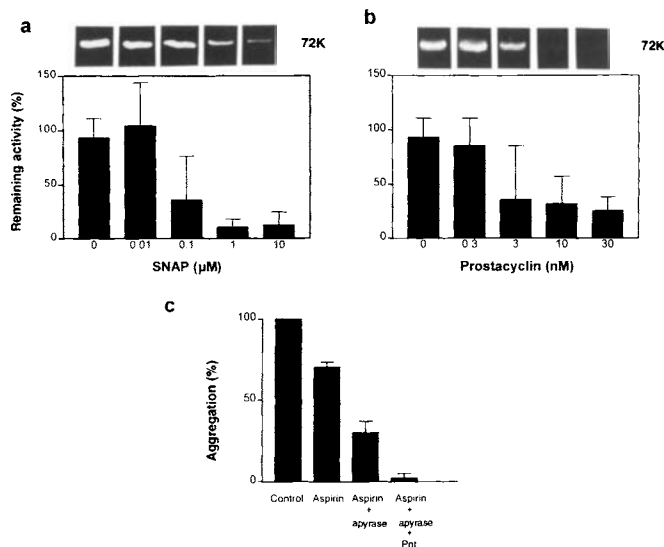


Figure 4 Inhibition of secretion of the 72K gelatinase by SNAP (**a**) and prostacyclin (**b**). Platelets were preincubated in the aggregometer before addition of SNAP or prostacyclin. Platelet aggregation and the release of pro-gelatinase A initiated by collagen (**a**) or thrombin (**b**) were assayed as described (Fig. 1). The effects of SNAP and prostacyclin on the release of pro-gelatinase A were calculated by measuring the remaining gelatinolytic activity relative to controls. Bars are mean $\pm$ s.d. from 5–7 experiments; the corresponding gels show data from a representative experiment. **c**, Inhibition of aspirin- and apyrase-insensitive aggregation by phenanthroline. The effects of aspirin (100 μ M), apyrase (250 μ g ml⁻¹) (both from Sigma) and phenanthroline (100 μ M) on aggregation induced by collagen (10 μ g ml⁻¹; control). Data are mean $\pm$ s.d., $n = 4$.

release from platelets. Recombinant gelatinase A does not cause aggregation of intact platelets but amplifies the pro-aggregatory effects of collagen, indicating that a cellular target for the action of this enzyme could be expressed upon platelet activation.

Gelatinase A at concentrations ≥ 100 ng ml⁻¹ (10^4 times higher than those that stimulate platelets) inhibited aggregation. These amounts of gelatinase A are likely to be expressed in diseased vasculature. There is an increased expression of these enzymes in vascular disorders such as atherosclerosis^{7,20–22}. It is likely that the growth and disruption of atherosclerotic plaques may depend, in part, on the amounts and activities of MMPs released during interactions between platelets, leukocytes and the vessel wall.

The discovery of a gelatinase-A-mediated pathway of human platelet aggregation and the aggregation-induced secretion of this enzyme suggest that gelatinase A may contribute to the physiological and pathological activity^{1–3,23,24} of platelets. As the inhibitors of MMPs abolish the aspirin-insensitive part of platelet aggregation, these compounds are good candidates for development as anti-thrombotic drugs that are more effective than aspirin. □

Methods

Blood was collected from healthy volunteers who had not taken any drugs for 2 weeks beforehand. Washed platelet suspensions (2.5×10^8 platelets per ml) were prepared as described²⁵. Samples (1.25×10^8 platelets in 0.5 ml Tyrode's solution) were centrifuged at 10,000g for 3 min at 4 °C and the supernatant (16 μ l) assayed for the presence of gelatinase A.

Gelatinase A and pro-gelatinase A were identified by zymography²⁶ and western blotting²⁷. Zymography was performed using 7% SDS-PAGE with copolymerized gelatin (2 mg ml⁻¹) as substrate²⁶. Enzyme activities were assayed using a ScanJet 3c scanner (Hewlett Packard) and SigmaGel measurement software (Jandel) with reference to the dose–response curves of the recombinant enzymes. Gelatinolytic activities were inhibited using phenanthroline (20 μ M; Sigma) and SC44463 (20 μ M; Pfizer).

As 72K gelatinase A is also found in blood leukocytes²⁸, their presence in washed platelet suspensions could generate the gelatinolytic activity; however, no leukocyte contamination of platelet suspensions was detected using a Coulter MicroDiff 16 whole-blood analyser ($<0.01\%$). In addition, zymography failed to detect the 92K gelatinase that is secreted with gelatinase A by leukocytes. It is thus unlikely that the release of gelatinase A from blood leukocytes could account for the enzyme detected in washed platelets.

Recombinant pro-gelatinase A was activated by incubation for 2 h at 22 °C in the presence of 2 mM 4-aminophenylmercuric acetate. For immunoprecipitation of gelatinase A, platelets were incubated overnight at 4 °C in the presence or absence of 1 μ g ml⁻¹ of anti-gelatinase A monoclonal antibodies (Oncogene Science). The resultant complex was precipitated using protein G–Sephacryl (Pharmacia Biotech), removed by centrifugation and assayed by zymography.

For western blotting, lysed platelets were subjected to 7% SDS–PAGE²⁷ and proteins identified by using anti-gelatinase A antibodies (1 μ g ml⁻¹; Oncogene Science).

For platelet aggregation, washed platelets (1.25×10^8 platelets in 0.5 ml Tyrode's solution) were preincubated in a whole-blood ionized-calcium lumiaggregometer (Chronolog) for 3 min at 37 °C. Platelet aggregation and ATP release²⁹ were initiated by collagen (0.5–5.0 μ g ml⁻¹; Chronolog), thrombin (0.01–0.30 U ml⁻¹; Chronolog) or phorbol myristate acetate (PMA; 0.032 μ M; Sigma) plus fibrinogen (200 μ g ml⁻¹; Sigma), maintained for 10 min and analysed using an Aggro-Link data-processing system. At the end of the incubation, platelet samples were centrifuged at 10,000g for 3 min at 4 °C and the supernatant tested for pro-gelatinase A and gelatinase A using zymography. In some experiments, inhibitors of aggregation or recombinant enzymes were preincubated in the aggregometer at 37 °C for 1–2 min before adding collagen, thrombin or PMA. The release of pro-gelatinase A and gelatinase A from platelets was calculated from standard curves for recombinant enzymes. Statistical analysis was done with GraphPad Prism software.

Received 12 December 1996; accepted 10 February 1997.

- Siess, W. Molecular mechanisms of platelet activation. *Pharmacol. Rev.* **69**, 58–178 (1989).
- Colman, R. W. Aggregin: a platelet ADP receptor that mediates activation. *FASEB J.* **4**, 1425–1435 (1990).
- Sargeant, P. & Sage, S. O. Calcium signaling in platelet and other non-excitable cells. *Pharmac. Ther.* **64**, 395–443 (1994).
- Ray, J. M. & Stetler-Stevenson, W. G. The role of matrix metalloproteinases and their inhibitors in tumour invasion, metastasis and angiogenesis. *Eur. Respir. J.* **7**, 2062–2072 (1994).
- Woessner, J. F. Jr in *Inhibition of Matrix Metalloproteinases: Therapeutic Approach* (eds Greenwald, R. A. & Golub, L. M.) Vol. 732, 11–21 (Ann. NY Acad. Sci., New York, 1994).
- Birkedal-Hansen, H. Proteolytic remodeling of extracellular matrix. *Curr. Opin. Cell Biol.* **7**, 728–735 (1995).
- Libby, P. Molecular bases of the acute coronary syndromes. *Circulation* **91**, 2844–2850 (1995).
- Petreguin, P. R., Tod, R. F., Devall, L. J., Boxer, L. A. & Currut, J. T. Association between gelatinase release and increased plasma membrane expression of the M01 glycoprotein. *Blood* **69**, 605–610 (1987).
- Seltzer, J. L. et al. Activation of 72-kDa type IV collagenase/gelatinase by normal fibroblasts in collagen lattices is mediated by integrin receptors but is not related to lattice contraction. *Exp. Cell Res.* **213**, 365–374 (1994).
- Salvesen, G. & Nagase, H. in *Proteolytic Enzymes: a Practical Approach* (eds Beynon, R. L. & Bond, J. S.) 83–104 (IRL, Oxford University Press, Oxford, New York and Tokyo, 1989).
- Cruywys, S. C., Davies, D. E. & Pettipher, E. R. Co-operation between interleukin-1 and the fibrinolytic system in the degradation of collagen by articular chondrocytes. *Br. J. Pharmacol.* **100**, 631–635 (1990).
- Sato, H. et al. A matrix metalloproteinase expressed on the surface of invasive tumour cells. *Nature* **370**, 61–65 (1994).
- Chesney, C. M., Harper, E. & Colman, R. W. Human platelet collagenase. *J. Clin. Invest.* **53**, 1647–1654 (1974).
- Leven, R. M. & Yee, T. Collagenase production by guinea pig megakaryocytes *in vitro*. *Exp. Hematol.* **18**, 743–747 (1990).
- Joseph, R., Welch, K. M. A., D'Andrea, G. & Riddle, J. M. Evidence for the presence of red and white cells within 'platelet' aggregates formed in whole-blood. *Thromb. Res.* **53**, 485–491 (1989).
- Stetler-Stevenson, W. G., Krutzsch, H. C. & Liotta, L. A. Tissue inhibitor of metalloproteinase (TIMP-2). A new member of the metalloproteinase inhibitor family. *J. Biol. Chem.* **264**, 17374–17379 (1989).
- Goldberg, G. I. et al. Human 72-kilodalton type IV collagenase forms a complex with a tissue inhibitor of metalloproteinases designated TIMP-2. *Proc. Natl. Acad. Sci. USA* **86**, 8207–8211 (1989).
- Vane, J. R. Inhibition of prostaglandin synthesis as a mechanism of action for aspirin-like drugs. *Nature (New Biol.)* **231**, 232–235 (1971).
- Komoszynski, M. & Wojtczak, A. Apyrases (ATP diphosphohydrolases, EC 3.6.1.5): function and relationship to ATPases. *Biochim. Biophys. Acta* **1310**, 233–241 (1996).
- Herron, G. S., Werb, Z., Dwer, K. & Banda, M. J. Secretion of metalloproteinases by stimulated capillary endothelial cells. I. Production of procollagenase and stromelysin exceeds expression of proteolytic activity. *J. Biol. Chem.* **261**, 2810–2813 (1986).
- Southgate, K. M., Davies, M., Booth, R. F. & Newby, A. C. Involvement of extracellular-matrix-degrading metalloproteinases in rabbit aortic smooth-muscle cell proliferation. *Biochem. J.* **288**, 93–99 (1992).
- Henney, A. M. et al. Localization of stromelysin gene expression in atherosclerotic plaques by *in situ* hybridization. *Proc. Natl. Acad. Sci. USA* **88**, 8154–8158 (1991).
- McEver, R. Leukocyte interactions mediated by selectins. *Thromb. Haemostasis* **66**, 80–87 (1991).
- Radomski, M. W. & Salas, E. in *Nitric Oxide: A Modulator of Cell–Cell Interactions in the Microcirculation* (ed. Kubas, P.) 43–74 (Springer, New York, 1995).

25. Radomski, M. W. & Moncada, S. An improved method for washing of human platelets with prostacyclin. *Thromb. Res.* **30**, 383–389 (1983).
26. Heussen, C. & Dowdle, E. B. Electrophoretic analysis of plasminogen activators in polyacrylamide gels containing sodium dodecyl sulfate and copolymerized substrates. *Analyt. Biochem.* **102**, 196–202 (1980).
27. Zembowicz, A., Jones, S. L. & Wu, K. W. Induction of cyclooxygenase-2 in human umbilical vein endothelial cells by lysophosphatidylcholine. *J. Clin. Invest.* **96**, 1688–1692 (1995).
28. Goetzl, E. J., Banda, M. J. & Leppert, D. Matrix metalloproteinases in immunity. *J. Immunol.* **156**, 1–4 (1996).
29. Feinman, R. D., Detwiler, T. C. & Ingberman-Wojenski, C. in *Platelets: Physiology and Pharmacology* (ed. Longenecker, G. L.) 429–440 (Academic, New York, 1985).

Acknowledgements. This work was supported, in part, by grants from the Alberta Heritage Foundation for Medical Research (AHFMR) and Eli Lilly Pharmaceuticals M.W.R. is an AHFMR Scholar. E.S. is an AHFMR–Eli Lilly postdoctoral fellow. We thank W. Steller-Stevenson (National Cancer Institute Laboratory of Pathology, NIH) and R. Pettipher (Pfizer Central Research) for gifts of human recombinant gelatinase A and SC44463, respectively. D. Rees (Glaxo-Wellcome) for prostacyclin and SNAP, and I. Martin (University of Alberta) for helpful comments.

Correspondence and requests for materials to M.W.R. (e-mail: Marek.Radomski@UAlberta.CA).

Caspase-1 processes IFN- γ -inducing factor and regulates LPS-induced IFN- γ production

Tariq Ghayur, Subhashis Banerjee, Margaret Hugunin, Deborah Butler, Linda Herzog, Adam Carter, Lucia Quintal, Les Sekut, Robert Talanian, Michael Paskind, Winnie Wong, Robert Kamen, Daniel Tracey & Hamish Allen

BASF Bioresearch Corporation, 100 Research Drive, Worcester, Massachusetts 01605-4314, USA

Interferon- γ -inducing factor (IGIF, interleukin-18) is a recently described cytokine that shares structural features with the interleukin-1 (IL-1) family of proteins and functional properties with IL-12^{1–4}. Like IL-12, IGIF is a potent inducer of interferon (IFN)- γ from T cells and natural killer cells^{1–3,5,6}. IGIF is synthesized as a biologically inactive precursor molecule (proIGIF). The cellular production of IL-1 β , a cytokine implicated in a variety of inflammatory diseases, requires cleavage of its precursor (proIL-1 β) at an Asp-X site by interleukin-1 β -converting enzyme^{7,8} (ICE, recently termed caspase-1⁹). The Asp-X sequence at the putative processing site in proIGIF^{2,3} suggests that a protease such as caspase-1 might be involved in the maturation of IGIF⁴. Here we demonstrate that caspase-1 processes proIGIF and proIL-1 β with equivalent efficiencies *in vitro*. A selective caspase-1 inhibitor blocks both lipopolysaccharide-induced IL-1 β and IFN- γ production from human mononuclear cells. Furthermore, caspase-1-deficient mice are defective in lipopolysaccharide-induced IFN- γ production. Our results thus implicate caspase-1 in the physiological production of IGIF and demonstrate that it plays a critical role in the regulation of multiple proinflammatory cytokines. Specific caspase-1 inhibitors would provide a new class of anti-inflammatory drugs with multipotent action.

The caspases are a family of intracellular cysteine proteases that display an absolute requirement for aspartic acid at the P1 position of their substrates⁹. ICE (caspase-1) plays a critical role in the production of IL-1 α and IL-1 β , as demonstrated with knock-out mice^{10,11}. CPP32 (caspase-3) plays a role in various apoptotic pathways^{12–14}. To test which caspase might accomplish the predicted Asp–Asn cleavage within the 192-residue murine proIGIF² molecule and the Asp–Tyr cleavage within the 193-residue human proIGIF³, we tested various recombinant human caspases (caspase-1, ICE; caspase-4, ICH-2; caspase-5, ICE rel_{III}; caspase-3, CPP32; caspase-6, Mch2; caspase-7, Mch3; and caspase-2, ICH-1)⁹ in an *in vitro* cleavage assay using radiolabelled, recombinant proIGIFs as substrates. The enzyme concentrations were normalized on the basis of cleavage of polyadenosyl ribosyl polymerase

Table 1 Proteolysis of PARP and proIGIF by recombinant caspases

Caspases	Concentration ($\mu\text{g ml}^{-1}$)	% Cleavage	
		PARP	ProIGIF
ICE (1)	1.25	99	100
ICH-2 (4)	5.00	93	82
ICE rel _{III} (5)	20.00	90	55
CPP32 (3)	5.00	100	0*
Mch2 (6)	10.00	96	2
Mch3 (7)	5.00	97	32†
ICH-1 (2)	75.00	98	5

PARP was used as a control substrate to demonstrate proteolytic capability of the different caspases¹². The concentrations of caspases and assay conditions used gave $\geq 90\%$ cleavage of 116K PARP to its 84K fragment. For proIGIF cleavage, levels of the 18K mature IGIF fragment were quantified and normalized using caspase-1 cleavage as 100%.

Caspases are numbered in parentheses as recommended in ref. 9.

* CPP32 ($5 \mu\text{g ml}^{-1}$) cleaved proIGIF but generated a 12K and a 10K fragment instead of the expected 18K fragment. Complete cleavage of PARP can be obtained with a 20-fold lower concentration of CPP32 ($0.25 \mu\text{g ml}^{-1}$, unpublished observations). CPP32 also cleaved murine proIGIF D35A mutant to generate the 12K and the 10K fragments. In contrast, the proIGIF D35A mutant is resistant to caspase-1 cleavage (Fig. 2).

† Unlike other caspases, the Mch3 precursor expressed in *E. coli* does not undergo autocatalysis to generate an active protease. Addition of caspase-1 was required to initiate Mch3 autocatalysis and generate an active Mch3 protease. Partial cleavage of proIGIF by Mch3 is mediated by the presence of caspase-1 in the Mch3 preparation (unpublished observations).

(PARP), a substrate common to all of the caspases. We found that only members of the caspase-1-related subfamily (caspases 1, 4 and 5)⁹ cleaved proIGIF to generate the expected 18K mature IGIF fragment (Table 1). ProIGIF was processed most efficiently by caspase-1, less efficiently by caspase-4, and partially by caspase-5. Before our studies, the major role of caspase-1 was believed to be the regulation of IL-1 β production through the processing of proIL-1 β ^{7,10,11}. Therefore, we compared the relative sensitivities to cleavage by human caspase-1 of human and murine proIL-1 β s with human and murine proIGIFs. We found that human caspase-1 processes proIGIF and proIL-1 β with the same efficiency *in vitro* (Fig. 1). Furthermore, the human and mouse proteins were equally sensitive to cleavage by human caspase-1.

To verify that caspase-1 cleaved murine proIGIF at the predicted Asp–Asn cleavage site², we generated a murine proIGIF mutant in which the aspartic acid residue at the predicted P1 position (D35) was substituted with alanine. The D35A mutant was not cleaved by caspase-1 (Fig. 2a). These results demonstrate that caspase-1 does cleave proIGIF at the predicted cleavage site. To determine if caspase-1-mediated cleavage of proIGIF is required for the generation of bioactive IGIF, we incubated murine proIGIF and proIGIF D35A proteins in the presence or absence of caspase-1 and tested for bioactivity (Fig. 2b) using induction of IFN- γ in murine splenocytes primed with concanavalin A (Con A)¹. The *in vitro* transcribed and translated proIGIF protein was biologically inactive as determined by its inability to induce IFN- γ production above the Con A control levels in splenocytes. This corroborates other data showing that the *Escherichia coli*-derived proIGIF protein is biologically inactive³. We detected biologically active IGIF following incubation of proIGIF with caspase-1; no appreciable IGIF bioactivity was detected with proIGIF D35A mutant control (Fig. 2b). In further experiments we showed that coexpression of murine proIGIF with either human caspase-1 or murine caspase-1 in COS cells results in the secretion of IGIF bioactivity (Table 2). Here the generation of active IGIF was caspase-1-dependent as no IGIF bioactivity was detected in supernatants of COS cells coexpressing either proIGIF D35A mutant and caspase-1, or proIGIF and an inactive human caspase-1 C285A mutant (Table 2). Together, our results suggest that proIGIF is another biologically relevant substrate for caspase-1.

To confirm that the cleavage of proIGIF by caspase-1, or an enzyme closely related to it, is required for IFN- γ induction in a cellular context, we tested the ability of caspase-1 inhibitors to block lipopolysaccharide (LPS)-induced IFN- γ production in human peripheral blood mononuclear cells (PBMCs). Fresh human PBMCs were stimulated *in vitro* with LPS in the presence or absence

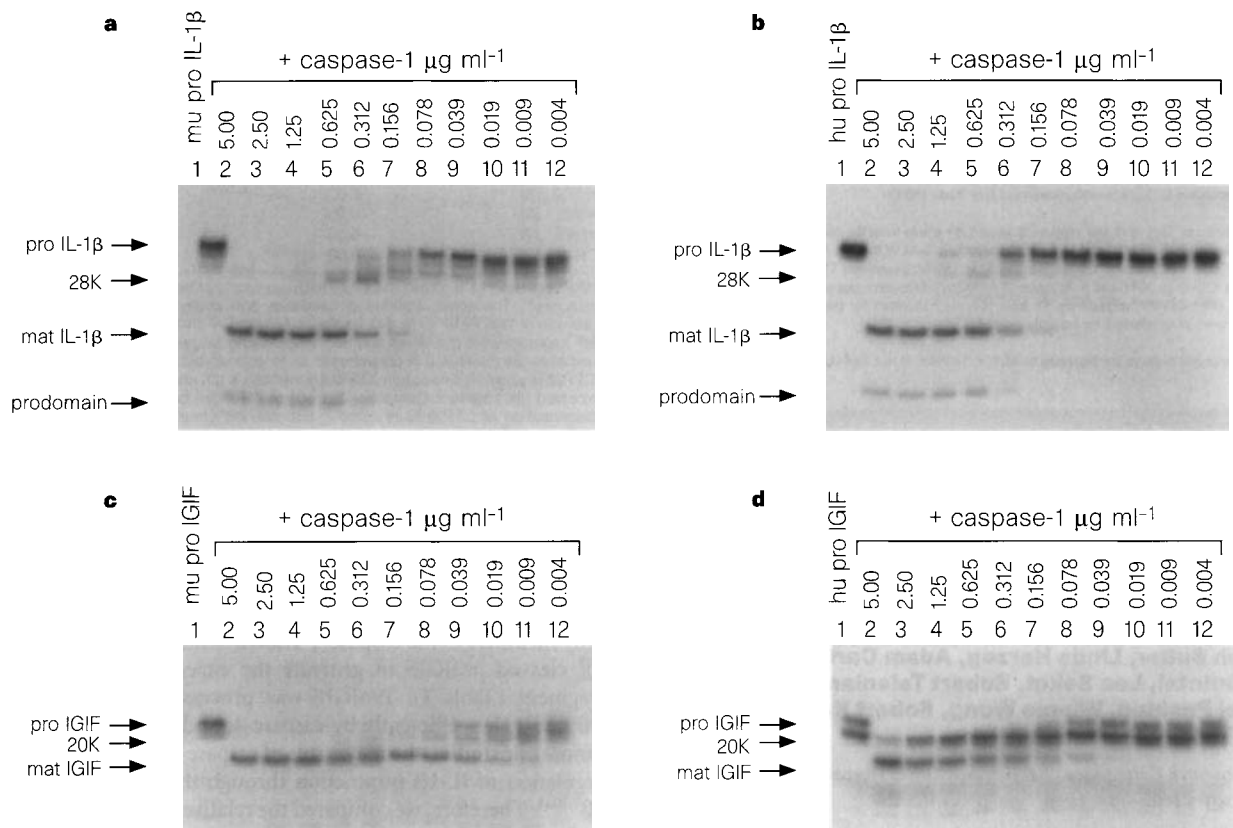


Figure 1 Cleavage of murine and human proIGIFs and proIL-1 β s by recombinant human caspase-1. The 20K band (**d**) is also derived from the human proIGIF cDNA. In this instance, the internal methionine residue at position 16 is probably

being utilized to initiate translation. This 20K band is cleaved by higher concentrations of human ICE to the 18K IGIF fragment. This 20K band was also observed with murine proIGIF, but to a lesser extent than for human proIGIF.

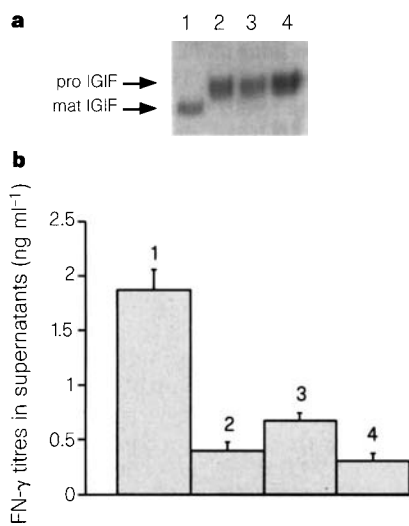


Figure 2 Cleavage of proIGIF by caspase-1 between Asp³⁵ and Asn³⁶ is required to generate bioactive IGIF. **a**, *In vitro* transcribed and translated murine proIGIF and proIGIF D35A mutant proteins incubated in the presence or absence of human caspase-1 (2.5 μ g ml⁻¹). Lane 1, proIGIF incubated with caspase-1; lane 2, proIGIF incubated without caspase-1; lane 3, proIGIF D35A mutant incubated with caspase-1; lane 4, proIGIF D35A mutant incubated without caspase-1. **b**, Bioactivity of *in vitro* transcribed and translated proIGIF and proIGIF D35A mutant following incubation with or without caspase-1 (for bar designation see **a**). The results are means $\pm$ s.d. of IFN- γ levels in Con A-stimulated murine splenocyte cultures from 2 experiments. IFN- γ levels for (1), proIGIF incubated with caspase-1, are significantly higher than levels in (2), (3) and (4) ($P < 0.05$). Splenocytes cultured in media alone released < 5 pg ml⁻¹ IFN- γ and splenocytes cultured with 1.25 μ g ml⁻¹ Con A released 193 ± 73 pg ml⁻¹ of IFN- γ in the supernatants.

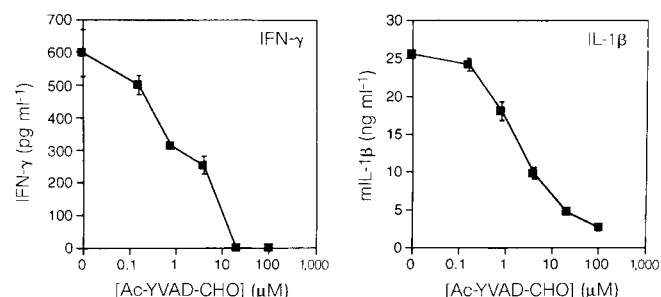


Figure 3 Caspase-1 inhibitor blocks LPS-induced IFN- γ and IL-1 β production by human PBMCs. Human PBMCs were stimulated with LPS (10 ng ml⁻¹, *E. coli* strain 0111:B4) in the presence or absence of the indicated concentrations of the caspase-1 inhibitor, Ac-YVAD-CHO (Bachem). Medium was collected and the levels of IFN- γ and mature IL-1 β were determined using ELISA kits from R&D Systems. The data shown are means $\pm$ s.d. for duplicate wells of PBMCs from a single donor. We have obtained similar results in five separate experiments with a total of five different donors: mean IFN- γ IC₅₀ = 2.3 ± 0.8 μ M and mean IL-1 β IC₅₀ = 2.4 ± 0.8 μ M.

of an inhibitor, Ac-YVAD-CHO⁷, that preferentially blocks the caspase-1 related caspases (caspase-1 50% inhibitory concentration (IC₅₀) = 17 nM and caspase-4 IC₅₀ = 200 nM¹⁵). This peptidic inhibitor blocked both IL-1 β and IFN- γ production by PBMCs with almost identical IC₅₀ values (IL-1 β IC₅₀ = 2.4 $\pm$ 0.8 μ M and IFN- γ IC₅₀ = 2.3 $\pm$ 0.8 μ M (n = 5 donors); Fig. 3). These observations suggest that caspase-1 is probably the dominant enzyme in PBMCs responsible for the generation of both LPS-induced IL-1 β and IGIF-mediated IFN- γ production.

Caspase-1-deficient (caspase-1^{-/-}) mice¹⁰ enabled us to evaluate the *in vivo* role of this enzyme in LPS-induced IFN- γ production. Administration of LPS induces high levels of serum IFN- γ , possibly through pathways dependent on IL-12 and/or IGIF^{2,3,16}. Caspase-1^{-/-} and wild-type (caspase-1^{+/+}) mice were injected with LPS and serum levels of IL-1 α , IL-1 β , IFN- γ and IL-12 heterodimer (p70) at various time points were determined by enzyme-linked immunosorbent assay (ELISA). Under these conditions, caspase-1^{-/-} mice had markedly lower IFN- γ levels, about 5% of the control levels in caspase-1^{+/+} mice 4–6 h after LPS injection (Fig. 4a). As reported previously¹⁰, caspase-1^{-/-} mice had almost undetectable levels of IL-1 α and IL-1 β (Fig. 4a). The drastic reduction in serum IFN- γ titres in caspase-1^{-/-} mice was observed despite normal serum levels of IL-12 heterodimer (p70) (Fig. 4b). Therefore, the reduction in LPS-induced IFN- γ production in caspase-1^{-/-} mice does not appear to be due to a defect in IL-12 production. IL-1 β has been implicated as a cofactor in the induction of IFN- γ by IL-12^{17,18}. As caspase-1^{-/-} mice are deficient in IL-1 β production, it is conceivable that the defect in IFN- γ production in caspase-1^{-/-}

mice is secondary to a deficiency in IL-1 β production. Administration of recombinant mouse IL-1 β *in vivo* did not significantly augment LPS-induced serum IFN- γ titres in either caspase-1^{+/+} or caspase-1^{-/-} mice (Fig. 5a). The bioactivity of the recombinant IL-1 β was confirmed by its ability to significantly increase serum IL-6 titres in both caspase-1^{+/+} and caspase-1^{-/-} mice (Fig. 5b). LPS-challenged caspase-1^{+/+} mice pretreated with neutralizing antibodies to IL-1 α and IL-1 β lacked detectable serum levels of IL-1 α and IL-1 β , but showed normal levels of serum IFN- γ (Fig. 6). Pretreatment with control IgGs did not influence LPS-induced serum IL-1 α , IL-1 β or IFN- γ levels in caspase-1^{+/+} mice. Thus, the defect in LPS-induced IFN- γ production in caspase-1^{-/-} mice does not result from lack of adequate levels of IL-1 α or IL-1 β . IGIF messenger RNA levels were similar in the spleens and livers of caspase-1^{+/+} and caspase-1^{-/-} mice (Fig. 6). We conclude from these studies that a defect in the caspase-1-dependent processing of proIGIF is the most likely explanation for the lack of LPS-induced IFN- γ production in caspase-1^{-/-} mice. Our finding that ICH-2 (caspase-4) can also process proIGIF offers an alternative hypothesis in which another caspase that requires caspase-1 for activation is actually responsible for IGIF processing *in vivo*, and further studies with additional knock-out strains will be necessary to exclude this more complex explanation rigorously.

The drastic reduction in LPS-induced serum IFN- γ titres in caspase-1^{-/-} mice is similar to that reported recently for LPS-stimulated IL-12-deficient mice¹⁶. Because both biologically active IGIF and IL-12 act synergistically to induce IFN- γ production *in vitro*^{2,6}, it is possible that both cytokines are required for LPS-

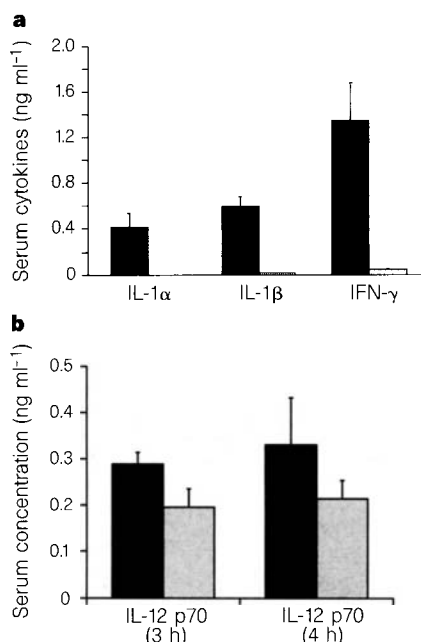


Figure 4 LPS-induced serum levels of IL-1 α , IL-1 β , IFN- γ and IL-12 (p70) in caspase-1^{-/-} and caspase-1^{+/+} mice. **a**, Solid bars, caspase-1^{+/+} mice; hatched bars, caspase-1^{-/-} mice. Serum levels were determined by ELISAs 4 h following LPS injection (40 mg kg⁻¹ i.v.)¹⁰. Similar defect in cytokine production was observed at 6 h after LPS challenge in caspase-1^{-/-} mice. The data shown are mean $\pm$ s.e.m. of 4 to 5 animals from a single experiment. These experiments were performed twice with similar results. Similar results were also obtained using mice injected with lower doses of LPS after presensitization with killed *Propionibacterium acnes* bacteria²⁴. Levels of IFN- γ at 4 h after LPS challenge were 166 $\pm$ 68 pg ml⁻¹ for caspase-1^{-/-} mice and 1293 $\pm$ 334 pg ml⁻¹ for caspase-1^{+/+} mice (P < 0.025). Murine IL-1 α and IL-1 β ELISA kits were from Endogen; IFN- γ ELISA kits were from Genzyme. **b**, Same designations and protocols as in **a**, except serum IL-12 p70 levels were determined 3 and 4 h following LPS injection. Murine IL-12 p70 ELISA kits were from Genzyme.

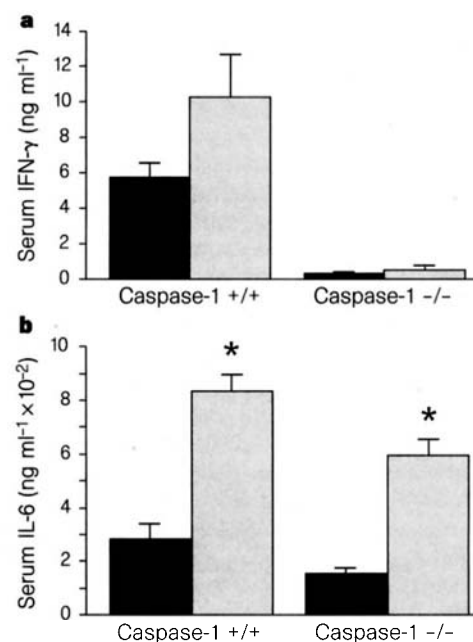


Figure 5 Effects of *in vivo* administration of recombinant mouse IL-1 β on LPS-induced serum IFN- γ and IL-6 responses in caspase-1^{+/+} and caspase-1^{-/-} mice. **a**, Serum IFN- γ titres; **b**, serum IL-6 titres. Solid bars, BSA treatment; hatched bars, recombinant mouse IL-1 β treatment (* P < 0.001). Results are expressed as mean $\pm$ s.e.m. (n = 10).

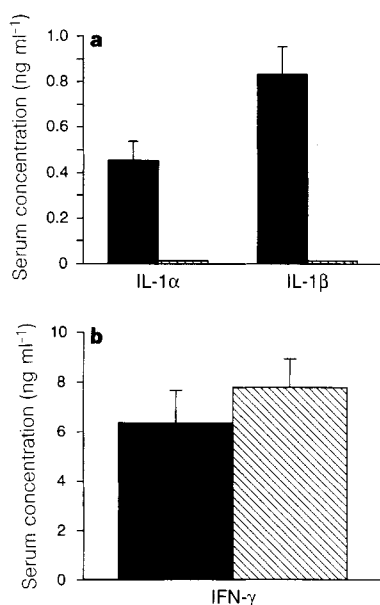


Figure 6 Effects of *in vivo* administration of anti-IL-1 α and anti-IL-1 β antibodies on LPS-induced serum IL-1 α , IL-1 β and IFN- γ responses in caspase-1^{-/-} mice. **a**, Serum IL-1 α and IL-1 β titres; **b**, serum IFN- γ titres. Solid bars, control IgG treatment; hatched bars, anti-IL-1 α + anti-IL-1 β IgG treatment. Results are expressed as mean $\pm$ s.e.m. ($n = 10$). Administration of antibodies. Caspase-1^{-/-} mice were treated i.p. with a mixture of mouse anti-IL-1 β antibody (250 μ g per mouse; Endogen) and goat anti-IL-1 α antibody (250 μ g per mouse; R&D), or a mixture of mouse IgG1 (250 μ g per mouse; Sigma) and goat IgG (250 μ g per mouse; Sigma) 30 min before injection with LPS (*E. coli* 0111:B4) at 40 mg kg⁻¹ i.v. All mice were bled 6 h after the LPS injection. Cytokine titres were determined by ELISA as described in the manuscript.

Table 2 Generation of bioactive IGIF following co-expression of murine proIGIF and murine or human caspase-1 in COS cells

Plasmids transfected	IFN- γ detected in supernatants (pg ml ⁻¹)
Medium alone	< 5
Con A*	181 $\pm$ 54
Con A + mock	415 $\pm$ 143
Con A + proIGIF	550 $\pm$ 84
Con A + proIGIF + hu caspase-1	1,908 $\pm$ 496*
Con A + proIGIF + C285A hu caspase-1	598 $\pm$ 141
Con A + proIGIF D35A	526 $\pm$ 190
Con A + proIGIF D35A + hu caspase-1	529 $\pm$ 161
Con A + proIGIF + mu caspase-1	2,292 $\pm$ 590†
Con A + proIGIF D35A + mu caspase-1	573 $\pm$ 46

Murine proIGIF or proIGIF D35A mutant were coexpressed in COS cells with murine caspase-1, human caspase-1 or an inactive human caspase-1 C285A mutant. Transfections were carried out using lipofectamine. Supernatants were collected 36 h post-transfection and tested for their ability to induce IFN- γ in Con A-primed murine splenocytes. IFN- γ titres were determined by ELISA (Genzyme). Results (mean $\pm$ s.d.) are from two separate experiments ($\dagger P < 0.05$).

* Con A 1.25 μ g ml⁻¹.

induced IFN- γ production *in vivo*. If this hypothesis is correct, then either a lack of production of bioactive IGIF (as occurs in caspase-1^{-/-} mice) or a lack of IL-12 (as in IL-12-deficient mice) would result in a dramatic reduction in LPS-induced IFN- γ production. IFN- γ is involved in the development and maturation of proinflammatory Th1 cells¹⁹. Thus, by regulating the production of IFN- γ (alone or in synergy with IL-12), IGIF may play a critical role in the pathogenesis of inflammatory processes in which IFN- γ is believed to be involved, such as colitis²⁰ and insulin-dependent diabetes mellitus²¹. Moreover, because the IFN- γ induced by LPS is probably derived from natural killer (NK) cells¹⁶, caspase-1 mediated maturation of IGIF probably plays a crucial role in regulating NK cell production of IFN- γ . NK cell-derived IFN- γ is believed to be

important in macrophage activation as part of an innate immune response²².

The results presented here extend the list of proinflammatory cytokines that are directly or indirectly regulated by caspase-1 to four: IL-1 β , IGIF, IL-1 α and IFN- γ . Caspase-1 therefore becomes an even more attractive target for the discovery of drugs to treat autoimmune inflammatory and other diseases in which these cytokines are implicated.

Since the submission of this manuscript, similar results have been published elsewhere²³.

Methods

cDNA cloning. Murine and human proIGIF cDNAs were cloned using the polymerase chain reaction (PCR) from an LPS-stimulated murine spleen cDNA library and LPS-stimulated THP-1 cells, respectively. PCR primers were designed on the basis of published proIGIF nucleotide sequences²⁴. Murine and human proIL-1 β s, and human PARP were cloned by reverse transcription PCR (RT-PCR), as described previously²⁵. The murine proIGIF D35A mutant was generated by an overlapping primer extension method using a WT murine proIGIF containing plasmid as a template. The overlapping internal PCR primer pair used was 5'-GACCTGGAATCAGCTAAGTTGGCCGACTTCAC-3' and 5'-GTGAAGTCGGCCAAAGTTAGCTGATTCCAGGTC-3'. Nucleotide sequences of all subcloned PCR fragments were confirmed before use.

In vitro cleavage assay. ³⁵S-methionine-labelled human and murine proIL-1 β s and proIGIFs, murine proIGIF D35A mutant, and PARP proteins were made using a TNT-coupled transcription and translation system (Promega)²⁵. Aliquots of *in vitro* translated, ³⁵S-labelled proteins were incubated with various concentrations of N-His-tagged, *E. coli*-derived caspases for 30 min at 25 °C. The cleavage reactions were stopped by addition of denaturing buffer and heating the samples. Cleavage profiles of various substrates were examined on 10–20% Tris-Tricine gradient polyacrylamide gels (Integrated Separation Systems) and autoradiography. For the quantification of proIGIF proteolysis by various caspases (Table 1), gels were dried and scanned (Biorad GS-250 Molecular Imager).

IGI bioassay. IGIF bioactivity was assessed using the protocol described by Okamura *et al.*². Spleen cells (0.8 $\times$ 10⁶ cells) were incubated with either *in vitro* translated IGIF (1:40 dilution, Fig. 2b) or supernatants derived from COS cells coexpressing murine proIGIF and murine or human caspase-1 (Table 2), plus 1.25 μ g ml⁻¹ Con A for 24 h at 37 °C in RPMI 1640 with 5% FCS, glutamine, penicillin and streptomycin. Following incubation, supernatants were collected and IFN- γ titres were determined by ELISA.

In vivo administration of IL-1 β . Mice were injected with 40 mg kg⁻¹ LPS (*E. coli* 0111:B4) intravenously and then treated intraperitoneally with either 0.1% BSA or 1 μ g of recombinant mouse IL-1 β (R&D Systems) 4 h after the LPS challenge. The mice were bled 6 h after LPS challenge and serum cytokine titres determined by ELISAs; IFN- γ ELISA was from Genzyme and IL-6 ELISA from Biosource.

Received 13 December 1996; accepted 25 February 1997.

- Nakamura, K. *et al.* Purification of a factor which provides a costimulatory signal for gamma interferon production. *Infect. Immun.* **61**, 64–70 (1993).
- Okamura, H. *et al.* Cloning of a new cytokine that induces IFN-gamma production by T cells. *Nature* **378**, 88–91 (1995).
- Ushio, S. *et al.* Cloning of the cDNA for human IFN-gamma-inducing factor, expression in *Escherichia coli*, and studies on the biologic activities of the protein. *J. Immunol.* **156**, 4274–4279 (1996).
- Bazan, J. F., Timans, J. C. & Kastelein, R. A. A newly defined interleukin-1? *Nature* **379**, 591 (1996).
- Okamura, H. *et al.* A novel costimulatory factor for gamma interferon induction found in the livers of mice causes endotoxic shock. *Infect. Immun.* **63**, 3966–3972 (1995).
- Micallef, M. J. *et al.* Interferon-gamma-inducing factor enhances T helper 1 cytokine production by stimulated human T cells: synergism with interleukin-12 for interferon-gamma production. *Eur. J. Immunol.* **26**, 1647–1651 (1996).
- Thornberry, N. A. *et al.* A novel heterodimeric cysteine protease is required for interleukin-1 beta processing in monocytes. *Nature* **356**, 768–774 (1992).
- Fraser, A. & Evan, G. A license to kill. *Cell* **85**, 781–784 (1996).
- Alnemri, E. S. *et al.* Human caspase-1/CED-3 protease nomenclature. *Cell* **87**, 171 (1996).
- Li, P. *et al.* Mice deficient in IL-1 beta-converting enzyme are defective in production of mature IL-1 beta and resistant to endotoxic shock. *Cell* **80**, 401–411 (1995).
- Kuida, K. *et al.* Altered cytokine export and apoptosis in mice deficient in interleukin-1 beta converting enzyme. *Science* **267**, 2000–2003 (1995).
- Martin, S. J. & Green, D. R. Protease activation during apoptosis: death by a thousand cuts? *Cell* **82**, 349–352 (1995).
- Patel, T., Gores, G. J. & Kaufmann, S. H. The role of proteases during apoptosis. *FASEB J.* **10**, 587–597 (1996).
- Kuida, K. *et al.* Decreased apoptosis in the brain and premature lethality in CPP32-deficient mice. *Nature* **384**, 368–372 (1996).

15. Kamens, J. *et al.* Identification and characterization of ICH-2, a novel member of the interleukin-1 beta-converting enzyme family of cysteine proteases. *J. Biol. Chem.* **270**, 15250–15256 (1995).
16. Magram, J. *et al.* IL-12-deficient mice are defective in IFN gamma production and type 1 cytokine responses. *Immunity* **4**, 471–481 (1996).
17. Hunter, C. A., Chizzonite, R. & Remington, J. S. IL-1 β is required for IL-12 to induce production of IFN- γ by NK cells: a role for IL-1 β in the T cell independent mechanism of resistance against intracellular pathogens. *J. Immunol.* **155**, 4347–4354 (1995).
18. D'Andrea, A. *et al.* Interleukin 10 (IL-10) inhibits human lymphocyte interferon- γ production by suppressing natural killer cell factor/IL-12 synthesis in accessory cells. *J. Exp. Med.* **178**, 1041–1048 (1993).
19. Abbas, A. K., Murphy, K. M. & Sher, A. Functional diversity of helper T lymphocytes. *Nature* **383**, 787–793 (1996).
20. Neurath, M. F. *et al.* Antibodies to interleukin 12 abrogate established experimental colitis in mice. *J. Exp. Med.* **182**, 1281–1290 (1995).
21. Campbell, I. L. *et al.* Essential role for interferon-gamma and interleukin-6 in autoimmune insulin-dependent diabetes in NOD/Wehi mice. *J. Clin. Invest.* **87**, 739–742 (1991).
22. Locksley, R. M. Interleukin 12 in host defense against microbial pathogens. *Proc. Natl Acad. Sci. USA* **90**, 5879–5880 (1993).
23. Gu, Y. *et al.* Activation of interferon- γ inducing factor mediated by interleukin-1 β converting enzyme. *Science* **275**, 206–209 (1997).
24. Fletcher, D. S. *et al.* A synthetic inhibitor of interleukin-1 beta converting enzyme prevents endotoxin-induced interleukin-1 beta production in vitro and in vivo. *J. Interferon Cytokine Res.* **15**, 243–248 (1995).
25. Hugunin, M. *et al.* Protease activity of in vitro transcribed and translated *Caenorhabditis elegans* cell death gene (*ced-3*) product. *J. Biol. Chem.* **271**, 3517–3522 (1996).

Acknowledgements. We thank J. Voss, P. Li and T. Barlozzari for helpful discussions. We appreciate the assistance of C. Hollins, L. McGuinness and G. Cunha for ELISA assays.

Correspondence and requests for materials should be addressed to T.G. (e-mail: ghayur@biova.dnet.basf-ag.de).

Genetic instability in colorectal cancers

C. Lengauer, K. W. Kinzler & B. Vogelstein

The Howard Hughes Medical Institute and The Johns Hopkins Oncology Center, Baltimore, Maryland 21231, USA

It has long been considered that genetic instability is an integral component of human neoplasia^{1–3}. In a small fraction of tumours, mismatch repair deficiency leads to a microsatellite instability at the nucleotide sequence level^{4,5}. In other tumours, an abnormal chromosome number (aneuploidy) has suggested an instability, but the nature and magnitude of the postulated instability is a matter of conjecture. We show here that colorectal tumours without microsatellite instability exhibit a striking defect in chromosome segregation, resulting in gains or losses in excess of 10^{-2} per chromosome per generation. This form of chromosomal instability reflected a continuing cellular defect that persisted throughout the lifetime of the tumour cell and was not simply related to chromosome number. While microsatellite instability is a recessive trait^{6,7}, chromosomal instability appeared to be dominant. These data indicate that persistent genetic instability may be critical for the development of all colorectal cancers, and that such instability can arise through two distinct pathways.

Aneuploidy occurs frequently in colorectal and other cancers³. There are many potential causes for this aneuploidy, and it is unknown whether aneuploidy is associated with a persistent defect in chromosome segregation. We therefore studied eight commonly used colorectal cancer cell lines, four of which were aneuploid and four of which were near diploid. To measure the rate of change in chromosome number, fluorescence *in situ* hybridization (FISH) of interphase cells was performed with a panel of centromeric probes⁸. Clones of each of the eight lines were generated and expanded through a defined number of generations before examination by FISH.

The eight lines could be clearly segregated into two groups based on these analyses (Table 1). Each of the aneuploid lines (HT29, SW480, SW837 and LoVo) showed a dramatic variation in chromosome content among the cells of individual clones. For example, over half of the cells of HT29 clone A gave a variant number of

signals of chromosome 7 (either a loss or gain compared to the modal number of 3; Figs 1a and 2). A similar variation affected each of the nine chromosomes examined (Fig. 2 and Table 1). Both losses and gains of chromosomes were observed, with, for example, 34% of cells demonstrating losses and 12% of cells demonstrating gains of chromosome 15 in HT29 clone A (Fig. 2). Evaluation of independent clones of HT29 cells confirmed the reproducibility of these observations (Table 1).

That chromosomal instability (CIN) was not the simple end-product of tumorigenesis was evident from the results obtained with the near-diploid tumour lines, each of which was known to have a different form of instability (MIN, resulting from mismatch repair (MMR) deficiency). For example, most of the cells in HCT116 clone A contained two copies of both chromosomes 7 and 18 (Fig. 1b) and the number of variant HCT116 cells was similar to the background number observed in normal lymphocytes (Table 1). Comparable results were found for chromosomes 1, 3, 4, 8, 12 and 15 in both HCT116 clone A and in independent HCT116 clones, as well as in clones of the other near-diploid lines examined (DLD1, SW48 and RKO; Fig. 2 and Table 1).

As the cells constituting each of the clones evaluated had gone through a similar number of generations and were grown under

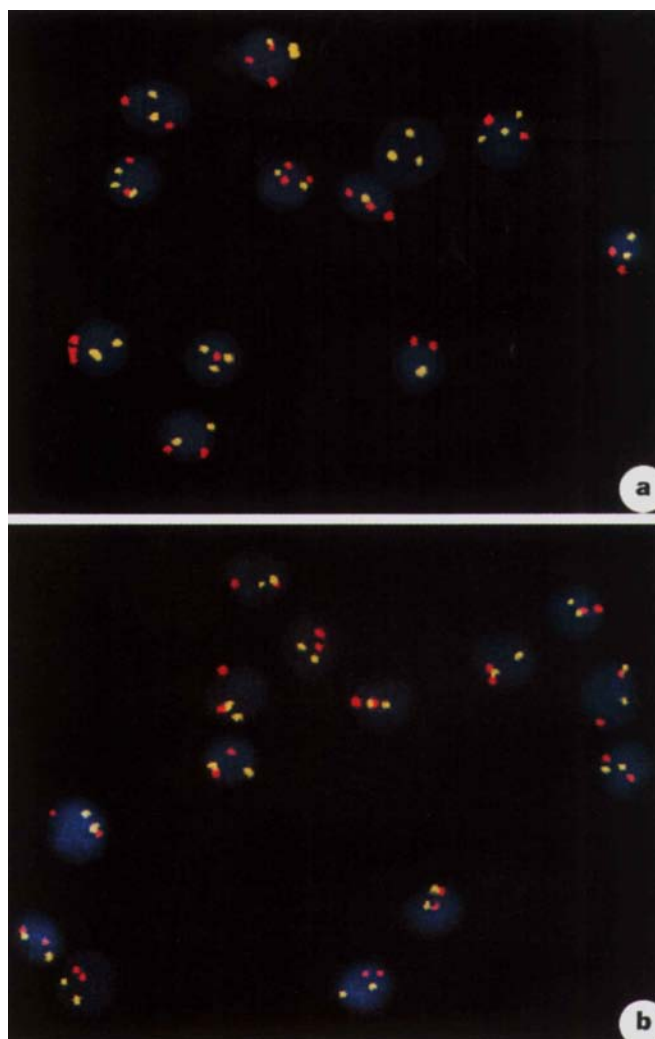


Figure 1 FISH analysis of colorectal cancer cell lines. Clones of HT29 (a) and HCT116 (b) were expanded through 25 generations before FISH analysis. Interphase nuclei were hybridized with labelled centromeric DNA probes specific for chromosome 7 (red) and 18 (yellow). HCT116 cells contained two copies of both chromosomes, whereas the number of signals detected in HT29 cells was diverse, indicating chromosomal instability.

Table 1 Chromosome instability in colorectal cancer cell lines

Cell line		Chrom 1	Chrom 3	Chrom 4	Chrom 7	Chrom 8	Chrom 12	Chrom 15	Chrom 16	Chrom 17	Chrom 18	Average
HT29	M	3	3	3	3	3	2	3	ND	3	3	2.9
Clone A	F	54%	53%	53%	55%	37%	56%	46%	ND	46%	59%	51%
HT29	M	3	3	3	3	3	2	3	ND	3	3	2.9
Clone A/50	F	56%	50%	48%	49%	48%	53%	60%	ND	45%	49%	51%
HT29	M	3	3	ND	3	2	3	3	2	3	3	2.8
Clone B	F	57%	48%	ND	53%	43%	30%	50%	54%	33%	49%	46%
LOVO	M	ND	4	4	6	4	4	2	ND	3	4	3.9
Clone A	F	ND	36%	25%	36%	26%	44%	25%	ND	9%	30%	29%
SW480	M	ND	5	4	7	6	6	7	4	5	4	5.3
Clone A	F	ND	59%	30%	57%	27%	41%	66%	34%	37%	34%	43%
SW837	M	2	2	ND	2	3	2	2	2	ND	1	2
Clone A	F	48%	20%	ND	19%	24%	16%	30%	33%	ND	16%	26%
SW837	M	2	2	ND	ND	3	2	2	ND	ND	ND	2.2
Clone B	F	17%	32%	ND	ND	35%	22%	28%	ND	ND	ND	27%
HCT116	M	2	2	2	2	2	2	2	ND	ND	2	2
Clone A	F	6%	5%	3%	3%	1%	2%	4%	ND	ND	4%	4%
HCT116	M	2	2	2	2	2	2	2	2	ND	2	2
Clone A/50	F	1%	7%	3%	4%	1%	1%	0%	3%	ND	9%	3%
HCT116	M	2	2	ND	2	2	2	2	ND	2	2	2
Clone B	F	3%	6%	ND	1%	4%	2%	5%	ND	1%	1%	3%
DLD1	M	2	2	ND	2	ND	2	2	ND	2	2	2
Clone A	F	2%	5%	ND	12%	ND	10%	2%	ND	2%	7%	6%
DLD1	M	ND	ND	ND	2	ND	2	ND	ND	ND	2	2
Clone B	F	ND	ND	ND	5%	ND	7%	ND	ND	ND	7%	6%
SW48	M	ND	2	ND	3	ND	2	ND	ND	2	2	2.2
Clone A	F	ND	8%	ND	2%	ND	0%	ND	ND	0%	6%	3%
RKO	M	2	2	ND	2	ND	2	ND	ND	2	2	2
Clone A	F	10%	9%	ND	7%	ND	8%	ND	ND	10%	10%	8%
Normal Lymphocytes	M	2	2	ND	2	2	2	2	2	2	2	2
	F	1%	1%	ND	3%	4%	2%	7%	2%	5%	2%	3%

Clones of several cell lines and fusions were expanded through 25 generations before interphase FISH analysis using chromosome-specific centromeric probes. In two cases, HT29 Clone A/50 and HCT116 Clone A/50, the clones were expanded through 50 generations before FISH analysis. For each chromosome tested, the modal chromosome number (*M*) was determined, and the fraction of cells whose chromosome number is different from the mode (*F*), is shown. At least two clones were evaluated for each line studied, with almost identical results in each case; representative results are shown. At least 100 nuclei were evaluated per clone with each chromosome probe. ND, not determined.

identical conditions *in vitro*, the rate of chromosome number change could be directly compared. The non-CIN lines passaged for 50 generations demonstrated far less variation than CIN cells passaged for 25 generations, ruling out trivial differences in generation number as the basis for the noted differences (Table 1). The rate of chromosome loss or gain was estimated to be in excess of 10^{-2} per generation for each chromosome analysed in each of the four CIN lines. Corrected for the number of chromosomes per cell and assuming that CIN affects all chromosomes equally, this translated into a gain or loss of a chromosome in at least one of every five cell divisions.

To confirm these estimates of chromosome instability, we developed a method to analyse chromosome variation *in situ* on single colonies containing only ~100 cells. Cell lines were plated at low density on microscope slides so that independent clones could be evaluated after 6–7 generations. FISH was performed following fixation of the colonies without disaggregating them. The cells within such HCT116 colonies showed little chromosome variation (0, 0 and 0.4% variant cells with respect to chromosomes 7, 12 and 17, respectively). HT29 colonies exhibited much greater variation, with 30, 34 and 18% of cells variant for chromosomes 7, 12 and 17, respectively. The degree of chromosome instability calculated from these data is consistent with the estimates generated from later stages of clonal growth (Table 1).

As already noted, all of the lines demonstrating CIN were aneuploid, containing a modal number of 40 (SW837), 87 (LoVo), 119 (SW480) and 71 (HT29) chromosomes, whereas the other four lines were near diploid, each containing 45–47 chromosomes. This observation is compatible with the idea that an underlying CIN caused the observed aneuploidy^{2,9}. However, it was

equally compatible with the hypothesis that a cell that becomes aneuploid for any reason may thereafter be chromosomally unstable (that is, aneuploidy could cause CIN)^{10,11}. To test this possibility, two experiments were done. First, chromosomal stability was tested in a derivative of HCT116 cells in which a single extra copy of chromosome 3 was introduced⁶. As shown in Fig. 2 and Table 2, this derivative was as stable as its parent with respect to all five chromosomes tested, including the extra chromosome 3. Second, chromosomally stable diploid lines were converted into grossly aneuploid ones by fusing drug-marked diploid clones to one another. FISH analysis demonstrated that each fusion line contained the expected number of copies of each chromosome (the sum of the parental numbers). FISH also revealed a distinct difference in chromosome stability between the artificially created aneuploid cells and the naturally occurring aneuploid lines described above. For example, clones derived from HCT116–HCT116 or DLD1–DLD1 fusions each contained a modal number of four chromosomes 7, 12 and 18. The variation in the number of these chromosomes among the cells was minimal, like their diploid parents (Fig. 2 and Table 2). As a technical control for these experiments, HT29–HT29 fusions were tested and found to demonstrate the expected high level of chromosomal instability (Fig. 2 and Table 2). Thus an abnormal number of chromosomes in a cell, in and of itself, could not account for the high level of chromosome instability found in the aneuploid colorectal cancers.

Cell fusion experiments also provided a unique opportunity to test whether CIN acted in a dominant or recessive fashion at the cellular level. Most fusions between CIN and non-CIN cells were non-viable, probably because of the existence of complementing tumour-suppressor genes (refs 7, 12, 13, and unpublished data). We

Table 2 Chromosome instability in colorectal cancer-cell derivatives resulting from fusion

Cell line		Chrom 1	Chrom 3	Chrom 7	Chrom 12	Chrom 15	Chrom 17	Chrom 18	Average
HCT116 + Ch3	M	ND	3	2	2	ND	2	2	2.2
Clone A	F	ND	6%	7%	3%	ND	2%	6%	5%
HCT116 × HCT116	M	ND	4	4	4	ND	4	4	4
Clone A	F	ND	0%	2%	0%	ND	0%	1%	1%
HCT116 × HCT116	M	ND	4	4	4	ND	ND	4	4
Clone B	F	ND	6%	4%	0%	ND	ND	4%	4%
HCT116 × HCT116	M	ND	4	4	4	ND	ND	4	4
Clone C	F	ND	9%	6%	4%	ND	ND	2%	5%
HCT116 × HCT116	M	ND	4	4	4	ND	ND	ND	4
Clone D	F	ND	0%	0%	0%	ND	ND	ND	0%
DLD1 × DLD1	M	ND	ND	4	4	ND	ND	4	4
Clone A	F	ND	ND	9%	8%	ND	ND	12%	9%
DLD1 × HCT116	M	ND	ND	4	4	ND	ND	4	4
Clone A	F	ND	ND	6%	7%	ND	ND	10%	8%
HT29 × HT29	M	ND	6	6	5	ND	5	6	5.6
Clone A	F	ND	58%	58%	58%	ND	62%	56%	58%
HT29 × HT29	M	ND	6	6	6	ND	5	6	5.6
Clone B	F	ND	54%	52%	58%	ND	64%	56%	57%
HT29 × HT29	M	ND	5	6	6	ND	5	6	5.6
Clone C	F	ND	70%	60%	52%	ND	60%	62%	61%
DLD1 × HT29	M	5	ND	5	4	5	5	ND	4.8
Clone A	F	48%	ND	58%	50%	56%	48%	ND	52%
DLD1 × HT29	M	4	ND	5	4	4	5	ND	4.4
Clone B	F	54%	ND	54%	42%	47%	52%	ND	50%
DLD1 × HT29	M	5	ND	5	5	5	5	ND	5
Clone C	F	14%	ND	46%	44%	44%	36%	ND	28%
DLD1 × HT29	M	4	ND	5	5	5	4	ND	4.6
Clone D	F	55%	ND	29%	49%	64%	52%	ND	50%
DLD1 × HT29	M	5	ND	5	4	5	4	ND	4.6
Clone E	F	42%	ND	34%	65%	55%	54%	ND	50%
DLD1 × HT29	M	5	ND	5	5	5	5	ND	5
Clone F	F	60%	ND	33%	38%	57%	44%	ND	46%
DLD1 × HT29	M	4	ND	5	5	5	5	ND	4.8
Clone G	F	28%	ND	46%	53%	53%	56%	ND	47%
DLD1 × HT29	M	4	ND	5	5	5	4	ND	4.6
Clone H	F	32%	ND	34%	34%	50%	54%	ND	41%

Drug-resistant clones were fused with one another as indicated. For example, HCT116 × HCT116 represents the fusion between two clones of HCT116 cells, one marked with geneticin resistance and the other with hygromycin resistance. The line marked HCT116 + Ch3 resulted from transfer of a normal human chromosome 3 to HCT116 cells⁶. This derivative was tested in the absence of selection for the introduced chromosome 3 (which contained a geneticin-resistance marker gene). Modal chromosome numbers (*M*) and variant fractions (*F*) were determined as for Table 1. At least two clones were evaluated for each fusion and the stability levels were found to be very similar among independent clones; representative results are shown. At least 100 nuclei were evaluated per clone. ND, not determined.

noticed, however, that both HT29 cells (a CIN line) and DLD1 cells (a MIN line) had similar tumour-suppressor gene alterations, including inactivating mutations of the APC and p53 genes^{14–16}. We therefore predicted that fusions between HT29 and DLD1 lines would be viable, and this turned out to be the case. The MIN phenotype of DLD1 was corrected by the introduction of the HT29 cell genome, as expected for a 'recessive' trait. In contrast, the CIN phenotype of HT29 persisted in the hybrid cells. In each of 9 such clones tested, chromosomal variation was high, with variant frequencies averaging 48% for five different chromosomes, similar to the HT29 parental cells (Table 2). Controls for these experiments were provided by fusions of non-CIN cells, including DLD1 cells, to one another; each of 13 independent clones were evaluated, and each remained chromosomally stable, despite the induced aneuploidy (examples are shown in Table 2). These results indicate that the CIN phenotype of HT29 cells acts dominantly at the cellular level, at least in the DLD1 background.

Our results strongly support the view that aneuploidy in cancers reflects an underlying chromosomal instability^{2,9}. Though aneuploidy is sometimes equated with instability, it is important to note that aneuploidy is a state, not a rate, and aneuploidy could in theory result from many factors other than a persistently elevated rate of chromosomal change. For example, an abnormal or heterogeneous

chromosome number in a cancer cell could result from a small number of catastrophic mitoses during tumour evolution in the absence of any persistent underlying defect in chromosomal segregation. Similarly, aneuploidy could result from the preferential outgrowth of cells with an abnormal chromosome number, selecting for gains of chromosomes containing activated oncogenes and for losses of chromosomes containing tumour-suppressor genes. Such selection could lead to aneuploidy in the absence of any change in the underlying rate of chromosomal gain or loss. Our results exclude these possibilities through the demonstration of genuine differences in the rates of chromosomal gain and loss in cancers with CIN, compared to that in normal cells or in cancer cells with MIN. They also demonstrate that, unlike the situation in yeast¹⁰, induced aneuploidy in human cells does not intrinsically lead to a chromosome instability.

It thus appears that all colorectal cancers may be genetically unstable, but that this instability can arise through one of two different pathways. Our results are consistent with karyotypic observations demonstrating that mismatch repair-deficient tumours, accounting for ~15% of colorectal cancers (CRC) are often diploid, whereas other CRC are generally aneuploid^{17,18}. Only one of the eight lines studied (LoVo) was both MIN and CIN, suggesting that either type of instability may be sufficient for driving

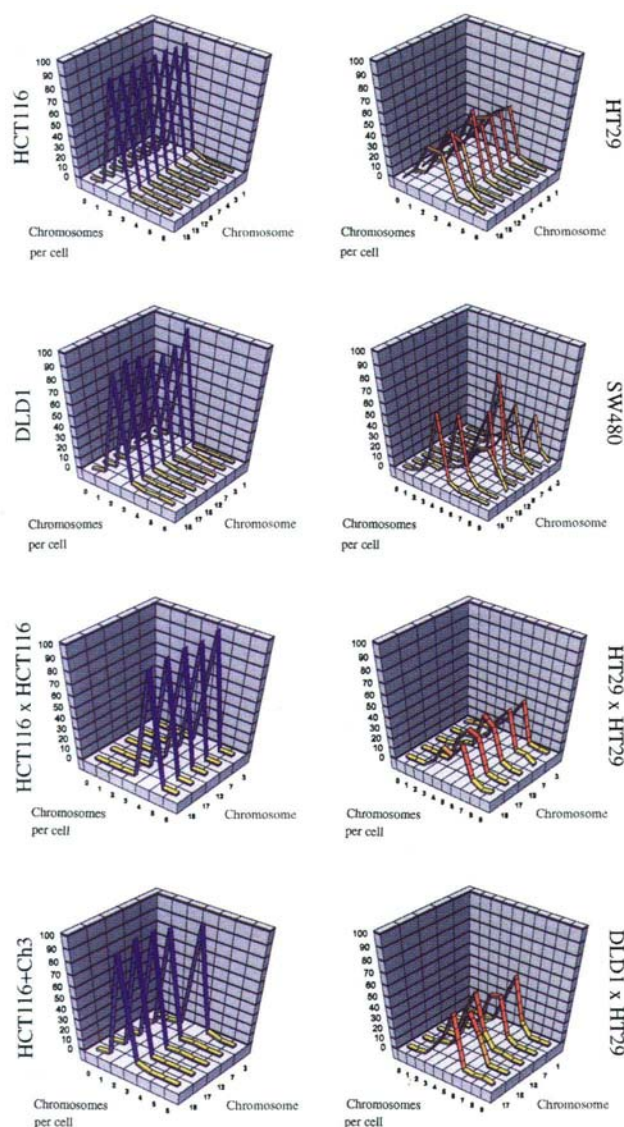


Figure 2 Ploidy and chromosomal instability. Clones of several cell lines and fusions were expanded through 25 generations, then analysed by FISH using chromosome specific centromeric probes. For each chromosome indicated, the fraction of cells with the indicated number of signals per nucleus is plotted along the z-axis. Blue ribbons represent fractions >90%; shades of red indicate fractions less than 75%. At least 100 nuclei were evaluated per clone for each chromosome tested.

the neoplastic process, but that MIN does not preclude CIN and vice versa. Although *p53* gene mutations have been invoked as a cause of some forms of genetic instability^{19,20}, such mutations are unlikely to be responsible for CIN, as there was no simple relationship between *p53* mutation and the CIN phenotype. For example, *p53* is mutant in the chromosomally stable DLD1 cells¹⁵, but is wild-type in the chromosomally unstable LoVo line²¹. We suggest that a gene affecting chromosome segregation is mutated in a dominant fashion early during colorectal tumorigenesis, and that the resultant instability drives the tumorigenic process in much the same way as MIN drives neoplasia in mismatch repair-deficient tumours. Further understanding of the genetic and biochemical mechanisms underlying this very common form of instability should prove important for understanding the tumour heterogeneity that is so central to cancer biology and treatment. □

Methods

Cell culture. Cells were cultured in McCoy's 5A medium supplemented with 10% fetal bovine serum (Hyclone Lab), 100 U ml⁻¹ penicillin, and 0.1 mg ml⁻¹ streptomycin. Monolayer cultures were grown at 37°C in a 5% CO₂ atmosphere. The cell lines DLD1, HCT116, HT29, LoVo, SW480 and SW837 were obtained from ATCC (Rockville, MD), and RKO was generously provided by M. Brattain. Their status with respect to microsatellite instability has been analysed previously^{15,22-24}. The cell lines DLD1, HT29, SW480 and SW837 contained mutant *p53* genes, whereas HCT116, LoVo, SW480 and RKO were *p53* wild-type (refs 14, 15, 21, 25, and our unpublished data). Cells were transfected with pCEP4 or pCEP9 using lipofectamine (Gibco BRL) and selected for resistance to geneticin or hygromycin, respectively. The resultant clones were evaluated for chromosomal instability between 25 and 50 generations following transfection.

Fluorescence in situ hybridization. Centromeric probes specific for chromosomes 1, 3, 4, 7, 8, 12, 15, 16, 17 and 18 were labelled with biotin-16-dUTP or digoxigenin-11-dUTP by nick translation⁸. Nuclei were fixed on slides and pretreated with RNase and pepsin as described²⁵. In some experiments (see text), cells were grown directly on microscopic slides and fixed in 3:1 methanol/acetic acid (vol/vol). Multicolour FISH was performed by following standard procedures²⁷. In most of the experiments, two differently labelled probes were hybridized and detected simultaneously. Biotinylated probes were detected with FITC avidin-DCS. Digoxigenin-labelled probes were detected with TRITC-conjugated rabbit-anti-mouse and goat-anti-rabbit antibodies. Cells were counterstained with DAPI. Evaluation of interphase nuclei was performed by standard epifluorescence microscopy (Zeiss Axiophot)²⁸. Photographs were taken with the 64 × objective using a CCD camera (Photometrics) equipped with a Kodak KAF 1400 chip. For digital image acquisition in the TIFF format, software package 'Nu200 2.0' (Photometrics) was applied. Further processing of images, which included only pseudocolouring and merging, was performed with the software package Gene Join²⁹.

Cell fusion. About 5 × 10⁶ cells from transfected, drug-resistant clones (geneticin or hygromycin) were combined and resuspended in 1 ml 50% PEG-1450 in McCoy's 5A medium and incubated at 37°C for 1 min, then washed in 20 ml McCoy's 5A. Cells were incubated overnight in growth medium, and drug selection (double selection with geneticin and hygromycin) started the following day. Doubly resistant fusion clones were isolated after 3 weeks.

Received 28 November 1996; accepted 11 February 1997.

- Loeb, L. Mutator phenotype may be required for multistage carcinogenesis. *Cancer Res.* **51**, 3075-3079 (1991).
- Hartwell, L. Defects in a cell cycle checkpoint may be responsible for the genomic instability of cancer cells. *Cell* **71**, 543-546 (1992).
- Heim, S. & Mitelman, F. *Cancer Cytogenetics* (Liss, New York, 1987).
- Marra, G. & Boland, C. R. Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. *J. Natl Cancer Inst.* **87**, 1114-1125 (1995).
- Bhattacharyya, N. P., Skandalis, A., Ganesh, A., Groden, J. & Meuth, M. Mutator phenotypes in human colorectal carcinoma cell lines. *Proc. Natl Acad. Sci. USA* **87**, 7555-7559 (1990).
- Koi, M. et al. Human chromosome 3 corrects mismatch repair deficiency and microsatellite instability and reduces N-methyl-N'-nitrosoguanidine tolerance in colon tumor cells with homozygous hMLH1 mutation. *Cancer Res.* **54**, 4302-4312 (1994).
- Casares, S., Ionov, Y., Ge, H.-Y., Standbridge, E. & Perucho, M. The microsatellite mutator phenotype of colon cancer cells is often recessive. *Oncogene* **11**, 2303-2310 (1995).
- Lichter, P., Boyle, A. L., Cremer, T. & Ward, D. Analysis of genes and chromosomes by nonisotopic in situ hybridization. *Gen. Anal. Tech. Appl.* **8**, 24-35 (1991).
- Hartwell, L., Weinert, T., Kadyk, L. & Garvik, B. Cell cycle checkpoints, genomic integrity, and cancer. *Cold Spring Harb. Symp. Quant. Biol.* **59**, 259-263 (1994).
- Mayer, V. W. & Aguilera, A. High levels of chromosome instability in polyploids of *Saccharomyces cerevisiae*. *Mut. Res.* **231**, 177-186 (1990).
- Shackney, S. et al. Model for the genetic evolution of human solid tumors. *Cancer Res.* **49**, 3344-3354 (1989).
- Tanaka, K. et al. Suppression of tumorigenicity in human colon carcinoma cells by introduction of normal chromosome 5 or 18. *Nature* **349**, 340-342 (1991).
- Goyette, M. C. et al. Progression of colorectal cancer is associated with multiple tumor suppressor gene defects but inhibition of tumorigenicity is accomplished by correction of any single defect via chromosome transfer. *Mol. Cell. Biol.* **12**, 1387-1395 (1992).
- Rodriguez, N. R. et al. *p53* mutations in colorectal cancer. *Proc. Natl Acad. Sci. USA* **87**, 7555-7559 (1990).
- Shibata, D., Peinado, M. A., Ionov, Y., Malkhosyan, S. & Perucho, M. Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nature Genet.* **6**, 273-281 (1994).
- Huang, J. et al. APC mutations in colorectal tumors with mismatch repair deficiency. *Proc. Natl Acad. Sci. USA* **93**, 9049-9054 (1996).
- Aaltonen, L. A. et al. Clues to the pathogenesis of familial colorectal cancer. *Science* **260**, 812-816 (1993).

18. Bocker, M. *et al.* Genomic instability in colorectal carcinomas: comparison of different evaluation methods and their biological significance. *J. Pathol.* **179**, 15–19 (1996).
19. Livingston, L. R. *et al.* Altered cell cycle arrest and gene amplification potential accompany loss of wild-type p53. *Cell* **70**, 923–935 (1992).
20. Yin, Y., Tainsky, M. A., Bischoff, F. Z., Strong, L. C. & Wahl, G. M. Wild-type p53 restores cell cycle control and inhibits gene amplification in cells with mutant p53 alleles. *Cell* **70**, 937–948 (1992).
21. Cottu, P. H. *et al.* Inverse correlation between RER⁺ status and p53 mutation in colorectal cancer cell lines. *Oncogene* **13**, 2727–2730 (1996).
22. Papadopoulos, N. *et al.* Mutation of the *mutL* homolog in hereditary colon cancer. *Science* **263**, 1625–1629 (1994).
23. Papadopoulos, N. *et al.* Mutations of GTBP in genetically unstable cells. *Science* **268**, 1915–1917 (1995).
24. Umar, A. *et al.* Defective mismatch repair in extracts of colorectal and endometrial cancer lines exhibiting microsatellite instability. *J. Biol. Chem.* **269**, 14367–14370 (1994).
25. Hollstein, M. *et al.* Database of p53 gene somatic mutations in human tumors and cell lines. *Nucleic Acids Res.* **22**, 3551–3555 (1994).
26. Ried, T. *et al.* Specific metaphase and interphase detection of the breakpoint region in 8q24 of Burkitt lymphoma cells by triple-color fluorescence *in situ* hybridization. *Genes Chrom. Cancer* **4**, 69–74 (1992).
27. Lichter, P. & Cremer, T. in *Human Cytogenetics: A Practical Approach* (eds Rooney, D. E. & Czepulkowski, B. H.) 157–192 (IRL, Oxford, 1992).
28. Lengauer, C. *et al.* Large-scale isolation of human 1p36-specific P1 clones and their use for fluorescence *in situ* hybridization. *Gen. Anal. Tech. Appl.* **11**, 140–147 (1994).
29. Ried, T., Baldini, A., Rand, T. C. & Ward, D. C. Simultaneous visualization of seven different DNA probes by *in situ* hybridization using combinatorial fluorescence and digital imaging microscopy. *Proc. Natl Acad. Sci. USA* **89**, 1388–1392 (1992).

Acknowledgements. This work was supported by the Fonds zur Förderung der wissenschaftlichen Forschung (Erwin-Schrödinger-Stipendium), the Clayton Fund, and by two grants from the NIH. B.V. is an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to B.V.

Activation of prokaryotic transcription through arbitrary protein–protein contacts

Simon L. Dove, J. Keith Joung* & Ann Hochschild

Department of Microbiology and Molecular Genetics, Harvard Medical School, 200 Longwood Avenue, Boston, Massachusetts 02115, USA

Many transcriptional activators in prokaryotes are known to bind near a promoter and contact RNA polymerase^{1–5}, but it is not clear whether a protein–protein contact between an activator and RNA polymerase is enough to activate gene transcription. Here we show that contact between a DNA-bound protein and a heterologous protein domain fused to RNA polymerase can elicit transcriptional activation; moreover, the strength of this engineered protein–protein interaction determines the amount of gene activation. Our results indicate that an arbitrary interaction between a DNA-bound protein and RNA polymerase can activate transcription. We also find that when the DNA-bound ‘activator’ makes contact with two different components of the polymerase, the effect of these two interactions on transcription is synergistic.

We replaced the carboxy-terminal domain (CTD) of the α -subunit of RNA polymerase (RNAP) (Fig. 1), which is the natural target for many transcriptional activators^{1–3}, with a heterologous protein domain that does not ordinarily mediate transcriptional activation. To do this, we took advantage of the well defined properties of the CTD of the bacteriophage λ cI protein (λ cI).

λ cI is a two-domain protein that binds DNA as a dimer, and pairs of dimers bind cooperatively to adjacent operator sites (Fig. 2a)⁵. The amino-terminal domain (λ cI-NTD) contacts the DNA and can interact with the σ subunit of RNAP, stimulating transcription when λ cI is bound at promoter λ P_{RM} (refs 6, 7). The λ cI-CTD mediates both dimer formation and the dimer–dimer interaction that results in cooperativity (reviewed in ref. 8) (Fig. 2a). λ cI

mutants specifically defective for transcriptional activation bear amino-acid substitutions in the λ cI-NTD⁹, and λ cI mutants specifically defective for cooperative binding to DNA bear amino-acid substitutions in the λ cI-CTD (ref. 10 and refs therein).

We reasoned that if we replaced the α -CTD with the λ cI-CTD, the resulting α -cI chimera would display a dimeric target that could be contacted by an appropriately positioned λ cI dimer (Fig. 2b). We wanted to investigate whether the same protein–protein interaction that ordinarily mediates the cooperative binding of pairs of λ cI dimers to the DNA would mediate transcriptional activation when the λ cI-CTD is tethered to the α -NTD.

We constructed a derivative of the *lac* promoter termed *plac* O_R2-62, bearing a single λ operator (O_R2) centred 62 base pairs (bp) upstream of the transcription startpoint (at –62) (Fig. 2b) and introduced it in single copy into the chromosome of *Escherichia coli* strain MC1000 F' *lacI*^q to create strain KS1. As expected, λ cI alone does not activate transcription from *plac* O_R2-62 (Fig. 3a), because when bound this far away from the promoter it cannot contact the σ -subunit of RNAP. However, λ cI stimulated transcription in the presence of the α -cI chimera approximately 10-fold, as measured by β -galactosidase assay (Fig. 3a). This stimulation was not dependent on the natural activating region located in the λ cI-NTD (data not shown). In the absence of the α -cI chimera, λ cI repressed transcription slightly. Furthermore, in the absence of λ cI, expression of the α -cI chimera had no significant effect on transcription from *plac* O_R2-62. Primer extension analysis confirmed that λ cI stimulated the production of correctly initiated transcripts (Fig. 3c).

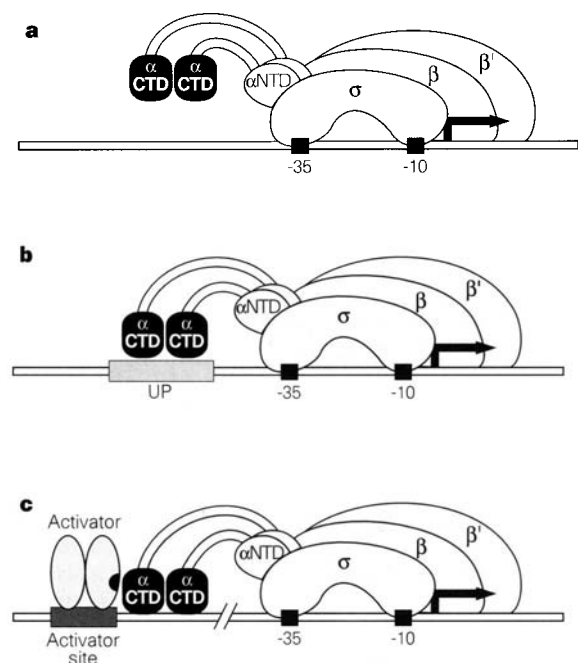


Figure 1 Function of the α -CTD in transcriptional activation. **a**, Basal transcription from a promoter that does not have an associated upstream element or transcriptional activator binding site. RNAP in *E. coli* consists of an enzymatic core composed of subunits α , β and β' in the stoichiometry $\alpha_2\beta\beta'$, and one of several alternative σ factors responsible for specific promoter recognition. **b**, Stimulated transcription from a promoter bearing an associated upstream element (UP) involves specific protein–DNA interaction between the α -CTD and the UP element. The α -CTD, which is tethered to α -NTD by a flexible linker region, can interact with the DNA UP element that is found upstream of the –35 region of certain particularly strong promoters³⁰. **c**, Stimulated transcription from a promoter bearing an associated activator binding site involves specific protein–protein interaction between the α -CTD and the DNA-bound activator.

* Present address: Massachusetts General Hospital, Department of Pathology, Division of Clinical Laboratories, Boston, Massachusetts 02114, USA.

The hypothesis that λ CI is activating transcription from *plac* O_R2-62 by contacting the λ CI-CTD fused to the α -subunit of RNAP predicts that a λ CI mutant unable to participate in the dimer-dimer interaction responsible for cooperativity would be unable to activate transcription in our artificial system. To test this prediction, we used a λ CI mutant (λ CI-D197G) that is unable to bind cooperatively to both adjacent and separated operator sites, but is indistinguishable from wild-type λ CI in its ability to dimerize and bind DNA¹⁰. Unlike wild-type λ CI, this mutant failed to activate transcription from *plac* O_R2-62 in the presence of the α -CI chimaera (Fig. 3b).

We then compared λ CI-D197G with two λ CI mutants with specific but less severe cooperativity defects¹⁰. Mutant N148D is less defective for cooperativity than mutant R196M, which is in turn less defective than D197G. Mutants N148D and R196M stimulated transcription from *plac* O_R2-62 more weakly than wild-type λ CI but more efficiently than λ CI-D197G; mutant R196M was the more defective of the two (Fig. 3b). Thus the strength of the interaction mediated by the λ CI-CTD correlated with the magnitude of the observed activation.

We have shown that λ CI bound at -62 can activate transcription by using its CTD to contact the λ CI-CTD fused to the α -subunit of RNAP. It has been previously shown that λ CI can also activate transcription when bound at -42 by using its NTD to contact the σ -subunit of RNAP⁵⁻⁷. We therefore wanted to see whether a single λ CI dimer bound at -42 (close to the promoter) could contact both the α -CI chimaera and the σ -subunit simultaneously, thereby activating transcription more efficiently than λ CI does ordinarily. To test this possibility, we used a P_{RM} -*lacZ* fusion bearing a single λ CI binding site (O_R2) located at its natural stimulatory position centred at -42. This λP_{RM} -*lacZ* fusion was introduced into *E. coli* strain MC1000 F'*lacI*^d on the low-copy plasmid vector $pP_{RM}\Delta$ -50 (ref. 11).

Whereas λ CI-stimulated transcription from P_{RM} resulted in ~130 units of activity in cells lacking the α -CI chimaera, it resulted in ~1,600 units in cells containing the α -CI chimaera, as measured by β -galactosidase assay (Fig. 4a). Primer extension analysis confirmed that λ CI stimulated the production of correctly initiated transcripts in both cases (Fig. 4c).

To show that the high level of λ CI-stimulated transcription observed in the presence of the α -CI chimaera depends on both activating surfaces of λ CI (the natural one located in its NTD and the artificial one located in its CTD), we made use of two λ CI mutants. One is specifically defective for activation through contact with the σ -subunit (λ CI-E34A)⁹, and the other is specifically defective for activation through contact with the α -CI chimaera (λ CI-D197G) (see above). Both mutants stimulated transcription much more weakly than wild-type λ CI (less than 10-fold), and, as expected, mutant E34A only stimulated transcription in the presence of the α -CI chimaera (Fig. 4a). We also verified that a double mutant bearing both substitutions failed to stimulate transcription. As when λ CI is bound further upstream, the strength of the interaction mediated by the λ CI-CTD correlated with the magnitude of the observed activation (Fig. 4b). We also demonstrated that substitution D197G, which abolishes λ CI cooperativity, has the same effect when it is present in λ CI, in the λ CI moiety of the α -CI chimaera, or in both (Fig. 4b).

The results with these λ CI mutants demonstrate that for an activator capable of contacting two components of RNAP, the effect of these two contacts on transcription can be synergistic. Specifically, the amount of transcription observed in the presence of wild-type λ CI and the α -CI chimaera (~1,600 units) is much greater than the sum of that observed in the presence of the two single mutants (80 and 140 units). Because the fractional occupancy of the λ CI binding site is ~85% under the conditions of the experiment shown in Fig. 4 (results not shown), this increased activation cannot be explained by an increase in the fractional occupancy of the λ CI binding site. Whereas the λ CI-CTD-mediated contact presumably

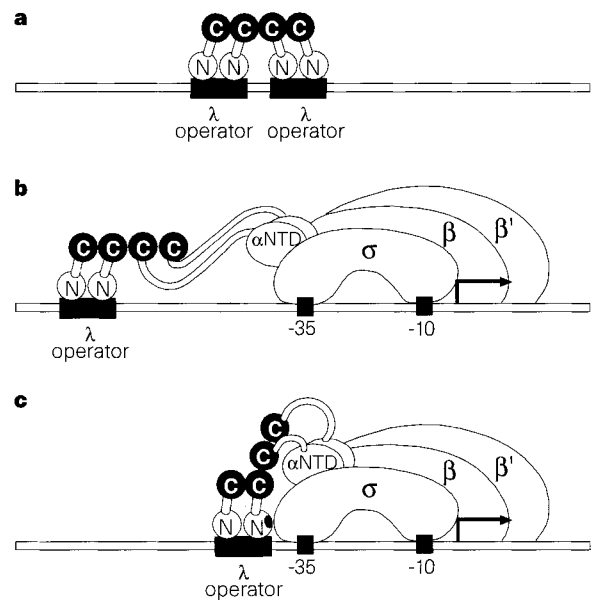


Figure 2 **a**, Interaction of adjacently bound λ CI dimers. Shown are two λ CI dimers (with each monomer depicted as a dumbbell to illustrate its domain structure) cooperatively bound to two adjacent operators (black boxes). The interaction between the two dimers, which is responsible for the cooperativity, is mediated by the CTD. **b**, Replacement of RNAP α -CTD by the λ CI-CTD permits interaction with the CTD of a DNA-bound λ CI dimer. The artificial promoter derivative *plac* O_R2-62 is shown: this bears the λ operator O_R2 centred 62 bp upstream of the transcriptional start site of the *lac* promoter. **c**, An appropriately positioned λ CI dimer can make two contacts with an RNAP molecule whose α -CTD has been replaced by the λ CI-CTD. The artificial promoter derivative $P_{RM}\Delta$ -50 is shown.

elicits activation simply by recruiting RNAP to the promoter, the λ CI-NTD-mediated contact has been shown to promote the formation of transcriptionally active 'open complexes'¹². Thus, in our artificial system, a single λ CI dimer is evidently able to exert both of these effects at once.

This artificial form of activation is like that of the *E. coli* cAMP receptor protein (CRP) at class II CRP-dependent promoters. When bound at a site centred near position -42, CRP makes two functional contacts with RNAP, using one activating region to contact the α -CTD and a second activating region to contact a target site in the α -NTD¹³.

Our results with two different promoters indicate that a protein domain can mediate transcriptional activation when tethered to the α -NTD simply by providing a surface that can be contacted by a DNA-bound protein. This activation presumably results from a strengthened association of RNAP with the promoter. Our findings are consistent with the proposal that natural activators that interact with the α -CTD function by recruiting RNAP to the promoter, but they indicate that a direct protein-DNA interaction involving the α -CTD (Fig. 1c) is not in principle essential for this domain to function as an activation target^{3,4}.

Our results suggest that a system analogous to the yeast two-hybrid system for detecting protein-protein interactions¹⁴ could now be used in *E. coli*. That is, our artificial system could in principle be used to detect any protein-protein interaction between a protein domain fused to the α -NTD and a second domain fused to a suitable DNA-binding domain. As the equilibrium dissociation constant for the interaction of λ CI dimers in solution is $\sim 10^{-6}$ M, and cooperative binding to DNA probably involves this same interaction¹⁵, any protein-protein interaction of comparable (or greater) strength could be expected to mediate transcriptional

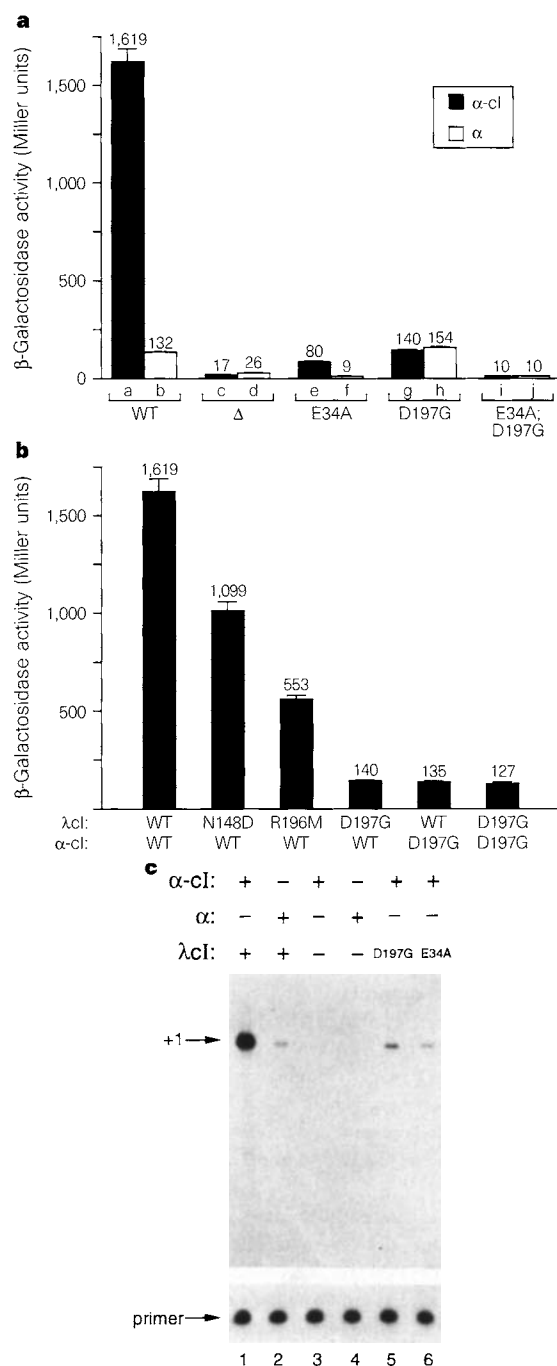
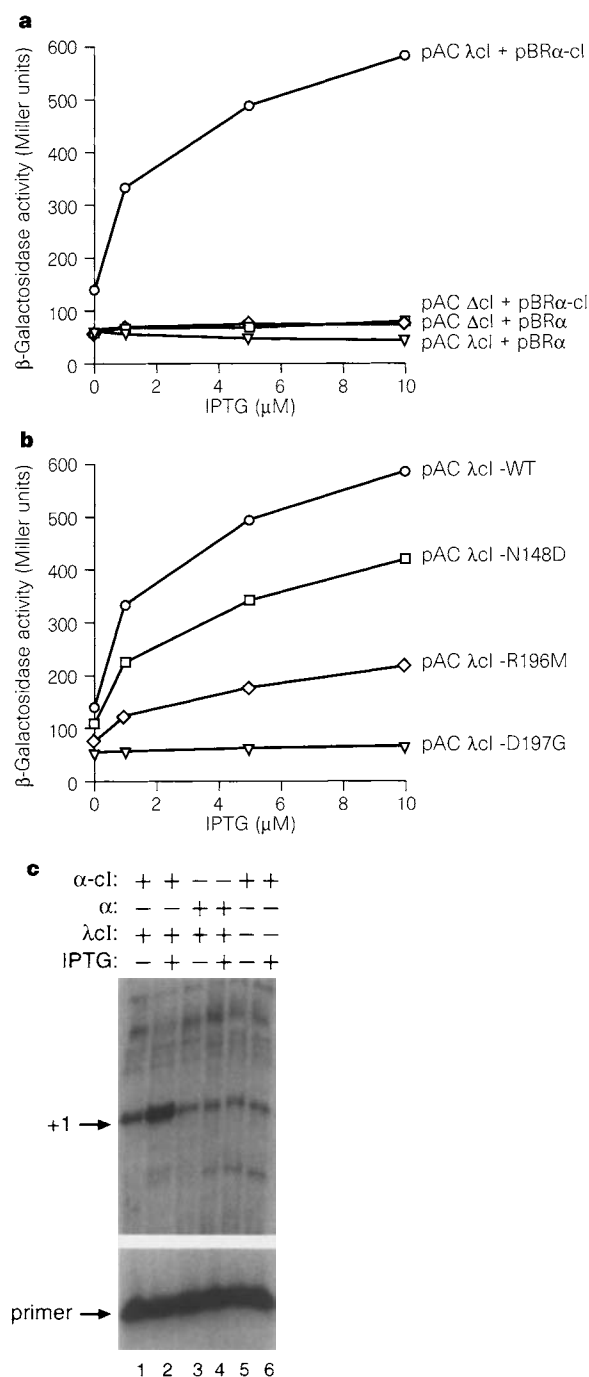


Figure 3 a, Effect of λcl on transcription *in vivo* from *plac* O_{R2-62} in the presence of α-cl fusion protein. KS1 cells harbouring the indicated plasmids were assayed for β-galactosidase as described²⁸. pACYC-derived plasmids either encoded λcl (pACλcl) or did not (pACΔcl); pBR322-derived plasmids either encoded the α-cl chimaera (pBRα-cl) or wild-type α (pBRα). **b**, Effects of λcl cooperativity mutants on transcription *in vivo* from *plac* O_{R2-62} in the presence of the α-cl chimaera. KS1 cells containing plasmid-encoded α-cl chimaera (pBRα-cl) together with plasmids expressing either wild-type λcl or the indicated cooperativity mutant derivative were assayed for β-galactosidase activity. The abilities of the mutants to bind cooperatively to adjacent and non-adjacent operators have been measured previously¹⁰: λcl wild type > λcl-N148D > λcl-R196M > λcl-D197G. **c**, Primer extension analysis of transcripts produced from *plac* O_{R2-62} in the presence of λcl and the α-cl chimaera. Total RNA was isolated from KS1 cells harbouring plasmids encoding the indicated proteins, and primer extension analysis was done by using a primer complementary to the *lacZ* transcript produced by the *plac* O_{R2-62} promoter. Primer extension products produced by correctly initiated *plac* O_{R2-62} transcripts are indicated by +1. Excess unincorporated primer is shown in the lower panel.

Figure 4 a, Effects of wild-type and mutant λcl proteins on transcription *in vivo* from P_{RMΔ-50} in the presence of α-cl chimaera or wild-type α. MC1000 F^{+/lacI} cells harbouring the reporter plasmid p_{RMΔ-50} and containing the indicated plasmid-encoded proteins were assayed for β-galactosidase. pACYC-derived plasmids directed expression of wild-type λcl (WT), no λcl (Δ), or the indicated λcl mutant, whereas pBR322-derived plasmids directed expression of either the α-cl chimaera (a, c, e, g, i) or wild-type α (b, d, f, h, j). **b**, Effects of λcl cooperativity mutants on transcription *in vivo* from P_{RMΔ-50} in the presence of α-cl chimaera. MC1000 F^{+/lacI} cells carrying the reporter plasmid p_{RMΔ-50} together with a pACYC-derived plasmid directing expression of either wild-type λcl (WT) or the indicated λcl cooperativity mutant, and a pBR322-derived plasmid directing expression of either the α-cl fusion protein (WT) or the D197G mutant derivative were assayed for β-galactosidase. **c**, Primer extension analysis of transcripts produced from P_{RMΔ-50} in the presence of both the α-cl chimaera and wild-type or mutant λcl protein. Total RNA was isolated from MC1000 F^{+/lacI} cells harbouring the reporter plasmid p_{RMΔ-50} together with plasmids encoding the proteins indicated. Primer extension analysis was done using the primer described previously¹. Primer extension products produced by correctly initiated P_{RM} transcripts are indicated by +1. Excess unincorporated primer is shown in the lower panel.

activation. We have shown that two interacting protein domains from yeast which interact more strongly than the λ CI-CTDs mediate greater transcriptional activation when fused respectively to the α -NTD and to λ CI (S.L.D., J.K.J. and A.H., manuscript in preparation). We do not know the interaction strength that would give maximal activation: depending on the particular promoter, and if the interaction between the DNA-binding protein and its recognition site were especially tight, a sufficiently strong protein-protein interaction might impede promoter clearance, resulting in repression¹⁶ rather than activation.

Our findings provide a parallel with recent findings in yeast^{17–21}. The discovery¹⁷ that transcriptional activation can be triggered by a fortuitous interaction between a DNA-bound protein lacking a classical activating region and a mutant form of a component of the RNAP II holoenzyme^{22,23}, suggested a simple recruitment model for gene activation¹⁷. (Recruitment of the TATA-binding protein to the promoter can also elicit transcriptional activation^{18–20}.) Analysis of the interaction involving a DNA-binding protein and a component of the RNAP II holoenzyme has established a correlation between its strength as measured *in vitro* and the degree of activation *in vivo*²¹. Further, addition of an acidic activating region to the DNA-binding protein, allowing additional contacts with the transcription machinery, results in a dramatic increase in the activation observed in the presence of the mutant holoenzyme²¹. These findings in yeast and ours in *E. coli* imply that activation in both prokaryotes and eukaryotes can be elicited by a simple protein-protein contact between a DNA-bound activator and an available target surface on the RNAP holoenzyme. □

Methods

Plasmids and strains. pAC λ CI harbours the wild-type λ CI gene under the control of the *lacUV5* promoter. Plasmid pKB280 (ref. 24) was cut with *EcoRI* and *HindIII* to yield a fragment containing the λ CI gene; the ends of this fragment, together with those of pACYC184 digested with *HindIII* and *HincII*, were made flush using Klenow and the two fragments ligated to generate pAC280. pAC λ CI was constructed by replacing the *HindIII*–*Bst* fragment of pAC280 with the corresponding *HindIII*–*Bam*HI fragment from pLR2 (ref. 10). pAC λ CI-D197G, pAC λ CI-R196M, pAC λ CI-N148D, pAC λ CI-E34A and pAC λ CI-E34A; D197G (made by cloning the appropriate restriction fragments from previously described plasmids^{10,11} into pAC λ CI) are identical to pAC λ CI except that they encode mutant derivatives of λ CI with the indicated amino-acid changes. pAC Δ CI is identical to pA3H Δ CI (ref. 10) and does not encode functional λ CI.

The plasmid pBR α -CI encodes residues 1–248 of the α -subunit of *E. coli* RNAP fused to residues 132–236 of λ CI under the control of tandem *lpp* and *lacUV5* promoters. The hybrid α -CI gene was amplified by the PCR and cloned into *EcoRI*–*Bam*HI-digested pBR α (a derivative of pHTF1 α ; ref. 25) to create pBR α -CI. The *HindIII*–*Bam*HI fragment from pLR2 (ref. 10) was cloned into pBR α -CI to make pBR α -CI. pBR α -CI-D197G was similarly constructed using the *HindIII*–*Bam*HI fragment from pLR2-D197G (ref. 10). PCR-amplified DNA was sequenced to verify that no errors had been introduced, and expression of α -CI fusion protein was confirmed by western blot using a polyclonal λ CI antiserum (data not shown).

The *lac* promoter derivative *plac* O_R2-62 was constructed by cleaving the plasmid KJ306 (ref. 26) with *HincII* and inserting a 31-bp linker sequence. *plac* O_R2-62 on a plasmid vector was transferred to a lysogenic phage and integrated in single copy into the chromosome of MC1000 *F'* *lacI*^q by lysogeny²⁷, creating strain KS1.

Experimental procedures. For the experiments shown in Fig. 3, cells were grown in LB supplemented with carbenicillin (50 μ g ml⁻¹), chloramphenicol (25 μ g ml⁻¹) and kanamycin (50 μ g ml⁻¹) together with IPTG at the concentration indicated. Assays were done at least three times in duplicate on separate occasions, with similar results. Values are the averages from one experiment; duplicate measurements differed by <10%. In the primer extension analysis (Fig. 3c), IPTG (isopropyl- β -D-thiogalactoside) was added to the growth

medium to a final concentration of 5 μ M where indicated. Conditions for experiments shown in Fig. 4 were as for Fig. 3, except that tetracycline (35 μ g ml⁻¹) was added to the growth medium. For Fig. 4, duplicate assays were done on duplicate cultures and data are expressed as the means of the four measurements. Experiments were done at least three times and a typical data set is shown; standard deviations were less than 10% and are indicated by error bars.

Primer extension analysis. RNA was isolated as described²⁸; primer labelling and primer extension assays were done essentially according to ref. 29. Transcriptional start sites were identified by electrophoresis of primer extension products alongside dideoxy sequencing reactions done with the same oligonucleotide used in the primer extension reactions (data not shown).

Received 9 December 1996; accepted 28 February 1997.

1. Ishihama, A. Protein-protein communication within the transcription apparatus. *J. Bacteriol.* **175**, 2483–2489 (1993).
2. Ishihama, A. Role of the RNA polymerase α subunit in transcription activation. *Mol. Microbiol.* **6**, 3283–3288 (1992).
3. Ebright, R. H. & Busby, S. The *E. coli* RNA polymerase α subunit: structure and function. *Curr. Opin. Genet. Dev.* **5**, 197–203 (1995).
4. Busby, S. & Ebright, R. H. Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell* **79**, 743–746 (1994).
5. Ptashne, M. *A Genetic Switch: Phage λ and Higher Organisms* (Cell Press and Blackwell Scientific, Cambridge, MA, 1992).
6. Li, M., Moyle, H. & Susskind, M. M. Target of the transcriptional activation function of phage λ CI protein. *Science* **263**, 75–77 (1994).
7. Kuldell, N. & Hochschild, A. Amino acid substitutions in the σ ⁻³⁵ recognition motif of σ ⁻³⁵ that result in defects in phage λ repressor-stimulated transcription. *J. Bacteriol.* **176**, 2991–2998 (1994).
8. Sauer, R. T., Jordan, S. R. & Pabo, C. O. λ repressor: a model system for understanding protein-DNA interactions and protein stability. *Adv. Protein Chem.* **40**, 1–61 (1990).
9. Bushman, F. D., Shang, C. & Ptashne, M. A single glutamic acid residue plays a key role in the transcriptional activation function of λ repressor. *Cell* **58**, 1163–1171 (1989).
10. Whipple, F. W., Kuldell, N. H., Cheatham, L. A. & Hochschild, A. Specificity determinants for the interaction of λ repressor and P22 repressor dimers. *Genes Dev.* **8**, 1212–1223 (1994).
11. Joung, I. K., Koepf, D. & Hochschild, A. Synergistic activation of transcription by bacteriophage λ CI protein and *E. coli* cAMP receptor protein. *Science* **265**, 1863–1866 (1994).
12. Hawley, D. K. & McClure, W. R. Mechanism of action of transcription initiation from the λ P_{RM} promoter. *J. Mol. Biol.* **157**, 493–525 (1982).
13. Niu, W., Kim, Y., Tan, G., Heyduk, T. & Ebright, R. H. Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. *Cell* **87**, 1123–1134 (1996).
14. Fields, S. & Song, O. A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245–246 (1989).
15. Sauer, R. T. *Molecular Characterization of the Lambda Repressor and its Gene ci*. PhD Thesis, Harvard Univ. (1979).
16. Monsalve, M., Mencia, M., Salas, M. & Rojo, F. Protein p4 represses phage ϕ 29 A2c promoter by interacting with the α subunit of *Bacillus subtilis* RNA polymerase. *Proc. Natl Acad. Sci. USA* **93**, 8913–8919 (1996).
17. Barberis, A. *et al.* Contact with a component of the polymerase II holoenzyme suffices for gene activation. *Cell* **81**, 359–368 (1995).
18. Chatterjee, S. & Struhl, K. Connecting a promoter-bound protein to TBP bypasses the need for a transcriptional activation domain. *Nature* **374**, 820–822 (1995).
19. Klages, N. & Strubin, M. Stimulation of RNA polymerase II transcription initiation by recruitment of TBP *in vivo*. *Nature* **374**, 822–823 (1995).
20. Xiao, H., Friesen, J. D. & Lis, J. T. Recruiting TATA-binding protein to a promoter: transcriptional activation without an upstream activator. *Mol. Cell Biol.* **15**, 5757–5761 (1995).
21. Farrell, S., Simkovich, N., Wu, Y., Barberis, A. & Ptashne, M. Gene activation by recruitment of the RNA polymerase II holoenzyme. *Genes Dev.* **10**, 2359–2367 (1996).
22. Koleske, A. & Young, R. A. An RNA polymerase II holoenzyme responsive to activators. *Nature* **368**, 466–469 (1994).
23. Kim, Y. J., Bjorklund, S., Li, Y., Sayre, M. H. & Kornberg, R. D. A multiprotein mediator of transcriptional activation and its interaction with the C-terminal repeat domain of RNA polymerase II. *Cell* **77**, 599–608 (1994).
24. Backman, K. & Ptashne, M. Maximizing gene expression on a plasmid using recombination *in vitro*. *Cell* **13**, 65–71 (1978).
25. Tang, H. *et al.* Location, structure, and function of the target of a transcription activator protein. *Genes Dev.* **8**, 3058–3067 (1994).
26. Joung, I. K. *Studies of Prokaryotic Transcriptional Activator Synergy and Dimerization Specificity*. PhD Thesis, Harvard Univ. (1996).
27. Hochschild, A. & Ptashne, M. Interaction at a distance between λ repressors disrupts gene activation. *Nature* **336**, 353–357 (1988).
28. Dove, S. L. & Dorman, C. J. The site-specific recombination system regulating expression of the type 1 fibrillar subunit gene of *Escherichia coli* is sensitive to changes in DNA supercoiling. *Mol. Microbiol.* **14**, 975–988 (1994).
29. Joung, I. K., Li, L. & Hochschild, A. Synergistic activation of transcription by *Escherichia coli* cAMP receptor protein. *Proc. Natl Acad. Sci. USA* **90**, 3083–3087 (1993).
30. Ross, W. *et al.* A third recognition element in bacterial promoters: DNA binding by the α subunit of RNA polymerase. *Science* **263**, 1407–1413 (1993).

Acknowledgements. We thank G. King for discussion, F. Whipple for plasmids, and M. Ptashne for comments on the manuscript. This work was supported by an NIH grant (A.H.) and by an established investigatorship from the American Heart Association (A.H.).

Correspondence and requests for materials should be addressed to A.H. (e-mail: ahochsch@warren.med.harvard.edu).